

Learning a Tree-Structured Ising Model in Order to Make Predictions

Guy Bresler[†] and Mina Karzand^{‡*}

Massachusetts Institute of Technology

Abstract: We study the problem of learning a tree graphical model from samples such that low-order marginals are accurate. We define a distance (“small set TV” or ssTV) between distributions P and Q by taking the maximum, over all subsets \mathcal{S} of a given size, of the total variation between the marginals of P and Q on \mathcal{S} . Approximating a distribution to within small ssTV allows making predictions based on partial observations. Focusing on pairwise marginals and tree-structured Ising models on p nodes with maximum edge strength β , we prove that $\max\{e^{2\beta} \log p, \eta^{-2} \log(p/\eta)\}$ i.i.d. samples suffices to get a distribution (from the same class) with ssTV at most η from the one generating the samples.

Keywords and phrases: Structure learning, Ising model, Inference, Graphical models, Trees.

1. Introduction

Markov random fields, or undirected graphical models, are a useful way to represent high-dimensional probability distributions. Their practical utility is in part due to: 1) edges between variables capture direct interaction, which make the model *interpretable* and 2) the graph structure facilitates efficient approximate *inference* from partial observations, for example using loopy belief propagation or variational methods. The inference task of interest to us in this paper is evaluation of conditional probabilities or marginals, e.g. $P(X_i = + | X_{\mathcal{S}} = x_{\mathcal{S}})$ or $P(x_{\mathcal{S}})$, having observed a subset of values $x_{\mathcal{S}}$ for a set of variables \mathcal{S} .

In applications it is often necessary to learn the model from data. It makes sense to measure accuracy of the learned model in a manner corresponding to the intended use. While in some applications it is of interest to learn the graph itself, in many machine learning problems the focus is on making predictions. In the literature, learning the graph is called *structure learning*; this problem has been studied extensively in recent years, see e.g. [20, 7, 1, 24, 19]. In this paper we consider the problem of learning a good model *for the purpose of performing subsequent inference*. This objective has been called “inferning” [15], and has received significantly less attention.

*The list of authors is in alphabetical order.

[†]Guy Bresler is with the Department of Electrical Engineering and Computer Science at MIT and a member of LIDS and IDSS. guy@mit.edu

[‡]Mina Karzand is with the Department of Electrical Engineering and Computer Science at MIT and a member of LIDS, IDSS and RLE. mkarzand@mit.edu

Tree-structured graphical models have been particularly well-studied. Aside from their theoretical appeal, there are two reasons for the practical utility of tree models: 1) Structure learning for trees can be accomplished with smaller sample and time complexity as compared to loopy graphs, and 2) Efficient exact inference (computation of marginals) is possible using belief propagation. Hence, in this paper we focus on tree-structured Ising models.

The main question we address is how many i.i.d. samples are required in order to guarantee that subsequent inference computations are accurate. Since computation of marginals on trees is easy, the crux of the task is in learning a model with marginals that are close to those of the original model. One of the take home messages is that learning for the purpose of making predictions requires dramatically fewer samples than is necessary for correctly recovering the underlying tree. The central technical challenge is that our analysis must also apply when it is impossible to learn the true tree, and this requires careful control of the sorts of errors that can occur.

With the goal of making predictions in mind, we introduce a new loss function to evaluate learning algorithms based on the accuracy of low-order marginals. The loss between true distribution P and learned distribution Q is defined to be

$$\mathcal{L}^{(k)}(P, Q) \triangleq \sup_{\mathcal{S}: |\mathcal{S}|=k} d_{\text{TV}}(P_{\mathcal{S}}, Q_{\mathcal{S}}),$$

where $P_{\mathcal{S}}$ denotes the marginal on set \mathcal{S} . As discussed in Section 3.2, small loss $\mathcal{L}^{(k)}$ guarantees accurate posterior distributions conditioned on sets of size $k-1$.

In this paper we restrict attention to tree-structured Ising models with no external field. For tree $\mathbf{T} = (\mathcal{V}, \mathcal{E})$ on p nodes and edge parameters $\alpha \leq |\theta_{ij}| \leq \beta$ for $(i, j) \in \mathcal{E}$, each configuration $x \in \{-1, +1\}^p$ is assigned probability

$$P(x) = \exp \left(\sum_{(i,j) \in \mathcal{E}} \theta_{ij} x_i x_j - \Phi(\theta) \right), \quad (1.1)$$

where $\Phi(\theta)$ is the normalizing constant.

Our main result gives lower and upper bounds on the number of samples needed to learn a tree Ising model to ensure small $\mathcal{L}^{(2)}$ loss, which in this setting is equivalent to accurate pairwise marginals. We emphasize that the task is to learn a model from the same class (tree-structured Ising) with these guarantees; this is sometimes called *proper* learning.

Theorem 1.1. *Given $n > C \max\{\eta^{-2} \log \frac{p}{\delta\eta}, e^{2\beta} \log \frac{p}{\delta}\}$ samples generated according to a tree Ising model P , denote the Chow-Liu tree by \mathbf{T}^{CL} . The Ising model Q on \mathbf{T}^{CL} obtained by matching correlations satisfies $\mathcal{L}^{(2)}(P, Q) \leq \eta$ with probability at least $1 - \delta$. Conversely, if $n \leq C' \eta^{-2} \log p$, then no algorithm can find a tree model Q such that $\mathcal{L}^{(2)}(P, Q) \leq \eta$ with probability greater than half.*

The result shows that the Chow-Liu tree, which can be found in time $O(p^2 \log p)$, gives small $\mathcal{L}^{(2)}$ error. We next place the result in context of related work.

1.1. Related work

Tree structured Markov random fields have attracted a lot of attention in different fields. These models, although admittedly a restricted class of probability distributions, are desirable from two different reasons: First, almost all computational tasks (e.g., computation of marginals or mode) that are hard even to approximate for general distributions can be performed exactly by efficient iterative algorithms on trees (typically in linear time). Sum-product or max-product are two well-studied examples [23, 32, 33, 17]. Second, learning tree-structured graphical models is also usually a much easier task in term of both sample and time complexity compared to the loopy graphs.

Structure learning in general graphical models have been studied widely. Information theoretic bounds on the number of samples have been provided [25, 29, 5, 6, 34]. Structure learning in trees and forest estimation have been studied by [8, 27]. Some generalizations of tree-structured models are studied such as forest approximations [28, 18], polytrees [12], bounded tree-width graphs [26, 22] and mixtures of trees [3, 19].

Structure learning becomes statistically more challenging, meaning more data is required, when interactions between variables are very weak or very strong [25, 5, 29]. It is intuitively clear that very weak edges are difficult to detect, leading to non-identifiability of the model. The goal of this paper is to show that accurate inference is possible even when structure learning is not.

Loopy belief propagation yields accurate marginals in high girth graphs (locally tree-like graphs) with correlation decay. [15] use this intuition to study this family of graphical models and propose an algorithm which recovers all the edges of the true graph. Due to insufficiency of the number of samples, the output of their algorithm can have extra edges, which are proved to be weak. The learned distribution is close to the true one in Kullback-Leibler divergence.

The paper [16] also studied recovery of the wrong graph for Gaussian graphical models and provided a lower bound on the KL distance between the estimated distribution and original distribution.

In [22], a polynomial time algorithm using polynomial number of samples is proposed to learn bounded tree-width graphical models with respect to KL divergence. They use ideas from submodular optimization and proper factorization of the distribution over bounded tree-width graphs.

Latent tree models have been well-studied in the phylogenetic reconstruction literature. The paper [14] studied sample and time complexity of tree metric based algorithms to reconstruct phylogenetic trees. [13] and [21] use distorted tree metrics to get approximations of the phylogenetic trees when the exact reconstruction of tree is not possible. In [21] studied in Section 6 a forest approximation of the tree is recovered. In [13] the prior assumptions on the phylogenetic tree is removed and instead a forest structure is recovered which is guaranteed to recover strong edges and deeper parts of the tree.

Similar approaches are taken in numerical taxonomy problems in which data is fitted to a tree metric. In [2], given a pairwise distance matrix D over p nodes is approximated by a tree metric with the induced distance matrix T .

Let $\epsilon = \min_{\mathbf{T}} \{\|T - D\|_{\infty}\}$. An $O(p^2)$ algorithm is provided which produces \hat{T} with $\|\hat{T} - D\|_{\infty} \leq 3\epsilon$ and it is shown that finding a T with $\|T - D\|_{\infty} \leq 9/8\epsilon$ is NP-hard.

For loopy graphs, both estimation of parameters and prediction based on noisy observations are computationally challenging tasks. [31] made the observation that considering the performance of the estimation and prediction together might be beneficial. They focus on computationally efficient prediction algorithms such as reweighted sum-product algorithm which is proved to converge to a unique global optima. They compare the prediction performance of these algorithms using two different sets of parameters. The first set is the true parameters of the model given by an oracle. The second set of parameters are derived by using the tree-reweighted Bethe surrogate. The second set of parameters are inconsistent in general and even having infinite data, this estimation process does not provide the true parameters. They show that using reweighted sum-product prediction algorithm, the inconsistent parameters derived by the tree-reweighted Bethe surrogate would provide smaller prediction error. It seems that the effect of error imposed in the estimation phase cancels out by using the approximate computationally efficient prediction algorithm.

1.2. Outline of paper

The next section provides background on the Ising model, tree models, and graphical model learning. Section 3.1 studies the problem of learning tree-structured Ising models, and specifies the sample complexity of exact recovery. Then in Section 3.2 we define a loss function motivated by inference computation and state our main result in Section 3.3. Section 4 analyzes an illustrative example to provide intuition for the main result. Section 5 contains a sketch of proof of the main result. Section 6 introduces the truncation algorithm and analyzes its performance in terms of ssTV. Section 7 contains proofs of the main results. A few technical lemmas are deferred to Section 9. Finally, Section 10 proves a combinatorial lemma on the relationship between the true and learned spanning trees.

2. Preliminaries

2.1. Notation

For a given tree $\mathbf{T} = (\mathcal{V}, \mathcal{E})$ let $\mathcal{P}_{\mathbf{T}}(\alpha, \beta)$ be the set of Ising models (1.1) with the restriction $\alpha \leq |\theta_{ij}| \leq \beta$ for each edge $(i, j) \in \mathcal{E}$. Denote by $\mathcal{P}_{\mathbf{T}} = \mathcal{P}_{\mathbf{T}}(0, \infty)$ the set of Ising models on \mathbf{T} with no restrictions on parameter strength. We denote by $\mu_{ij} = \mathbb{E}_P X_i X_j$ the correlation between the variables corresponding to any pair of vertices $i, j \in \mathcal{V}$. For an edge $e = (i, j)$ we write $\mu_e = \mu_{ij}$ and similarly for a set of edges $\mathcal{A} \subseteq \mathcal{E}$, $\mu_{\mathcal{A}} = \prod_{e \in \mathcal{A}} \mu_e$. The empirical distribution of samples is denoted by $\hat{P}(x) = \frac{1}{n} \sum_{l=1}^n \mathbf{1}_{\{X^{(l)}=x\}}$ and $\hat{\mu}_{ij} = \mathbb{E}_{\hat{P}} X_i X_j$ is the empirical correlation between nodes i and j .

2.2. Tree models

We study distributions that are Markov with respect to graphs. A distribution P on $\mathcal{X}^{\mathcal{V}}$ is Markov with respect to a graph $G = (\mathcal{V}, \mathcal{E})$ if for all $i \in \mathcal{V}$, we have $P(x_i | x_{\mathcal{V} \setminus \{i\}}) = P(x_i | x_{\partial i})$ (here ∂i is the neighborhood of i in G). In this paper we are interested in distributions P that are Markov with respect to a tree $\mathsf{T} = (\mathcal{V}, \mathcal{E})$. A consequence (see [17]) is that $P(x)$ factorizes as

$$P(x) = \prod_{i \in \mathcal{V}} P(x_i) \prod_{(i,j) \in \mathcal{E}} \frac{P(x_i, x_j)}{P(x_i)P(x_j)}. \quad (2.1)$$

For an arbitrary distribution Q and tree T , the distribution

$$\tilde{Q}(x) = \arg \min_{P \text{ is factorized as (2.1)}} D(Q \| P)$$

is the best approximation to Q within distributions Markov with respect to T . It is proved in [8] that \tilde{Q} is specified by matching the first and second moments, i.e., for all $(i, j) \in \mathcal{E}$, and all $x_i, x_j \in \mathcal{X}$, $\tilde{Q}(x_i, x_j) = Q(x_i, x_j)$.

For distribution P on $\{-, +\}^p$ and tree T , we define $\Pi_{\mathsf{T}}(P) = \arg \min_{Q \in \mathcal{P}_{\mathsf{T}}} D(P \| Q)$ to be the projection of P onto the class of Ising models on T with no external field. Hence, $\tilde{P} = \Pi_{\mathsf{T}}(P)$ can be represented as Equation (1.1) for some $\tilde{\theta}$ supported on T . Lemma 11.1 shows that $\tilde{P} = \Pi_{\mathsf{T}}(P)$ has edge weights $\tilde{\theta}_{ij}$ for each $(i, j) \in \mathcal{E}_{\mathsf{T}}$ satisfying $\tanh \tilde{\theta}_{ij} = \mu_{ij} \triangleq \mathbb{E}_P X_i X_j$.

3. Learning trees to make predictions

In this section we first state a specialization to our setting of the Chow-Liu algorithm. Later, in order to set up the learning for prediction problem, we discuss the problem of exact structure learning and give tight (up to a multiplicative constant) sample complexity for that problem. Finally, we define the ssTV distance $\mathcal{L}^{(k)}$, explain how it relates to prediction, and then state our results.

Denote the set of all trees on p nodes by \mathcal{T} . For some distribution $P \in \mathcal{P}_{\mathsf{T}}$, one observes n independent samples (configurations) $X^{(1)}, \dots, X^{(n)} \in \{-, +\}^p$ from the Ising model (1.1). Restricting to trees, a *structure learning algorithm* is a (possibly randomized) map $\phi : \{-1, +1\}^{p \times n} \rightarrow \mathcal{T}$ taking n samples $X^{1:n} = X^{(1)}, \dots, X^{(n)}$ to a tree $\phi(X^{1:n})$.

The maximum likelihood tree or Chow-Liu tree plays a central role in tree structure learning. Chow and Liu [8] observed that the maximum likelihood tree is the max-weight spanning tree in the complete graph, where each edge has weight equal to the empirical mutual information between the variables at its endpoints. The tree can thus be found greedily, and the run-time is dominated by computing empirical mutual information between all pairs of nodes.

Lemma 11.2 applies analysis similar to [8] to zero-field Ising models on trees to support the following definition:

Definition 3.1 (Chow-Liu Tree). Given n i.i.d samples $X^{1:n}$ from distribution $P \in \mathcal{P}_{\mathcal{T}}$, we define the Chow-Liu tree to be the maximum likelihood tree:

$$\mathsf{T}^{\text{CL}} = \operatorname{argmax}_{\mathsf{T} \in \mathcal{T}} \max_{P \in \mathcal{P}_{\mathsf{T}}} P(X^{1:n}).$$

This definition is slightly abusing the conventional terminology, as the Chow-Liu tree is classically the maximum likelihood tree assuming that the generative distribution is tree-structured, whereas in our definition we are assuming that the original distribution $P \in \mathcal{P}_{\mathcal{T}}$ can be described by Equation (1.1). Thus, it is not only tree-structured, but also has uniform singleton marginals.

Note that maximizing the likelihood of i.i.d. samples corresponds to minimizing the KL divergence. For the samples with empirical distribution \hat{P} , $\mathsf{T}^{\text{CL}} = \operatorname{argmin}_{\mathsf{T} \in \mathcal{T}} \min_{P \in \mathcal{P}_{\mathsf{T}}} D(\hat{P} \| P)$. Lemma 11.2 shows that

$$\mathsf{T}^{\text{CL}} = \operatorname{argmax}_{\{\text{spanning trees } \mathsf{T}'\}} \sum_{e \in \mathcal{E}_{\mathsf{T}'}} |\hat{\mu}_e|, \quad (3.1)$$

where for $e = (i, j)$, $\hat{\mu}_e = \mathbb{E}_{\hat{P}} X_i X_j$ is the empirical correlation between variables X_i and X_j .

Chow and Wagner [9] showed that the maximum likelihood tree is consistent for structure learning of general discrete tree models, i.e., in the limit of large sample size the correct graph structure is found. More recently, detailed analysis of error exponents was carried out by Tan et al. [27, 28]. A variety of other results and generalizations have appeared, including for example Liu et al.'s work on forest estimation with non-parametric potentials [18].

3.1. Exact recovery of trees

The statistical performance of a structure learning algorithm is often measured using the zero-one loss,

$$\mathcal{L}^{0-1}(\mathsf{T}, \mathsf{T}') = \mathbf{1}_{\{\mathsf{T} \neq \mathsf{T}'\}}, \quad (3.2)$$

meaning that the exact underlying graph must be learned. (See e.g., [7], [25], [28], [18]). The risk, or expected loss, under some distribution $P \in \mathcal{P}_{\mathcal{T}}(\alpha, \beta)$ is then given by the probability of reconstruction error, $\mathbb{E}_P \mathcal{L}^{0-1}(\mathsf{T}, \phi(X^{1:n})) = \mathbb{P}(\phi(X^{1:n}) \neq \mathsf{T})$, and for given α, β, p and n , the maximum risk is $\sup\{\mathbb{P}(\phi(X^{1:n}) \neq \mathsf{T}) : \mathsf{T} \in \mathcal{T}, P \in \mathcal{P}_{\mathcal{T}}(\alpha, \beta)\}$.

The sample complexity for learning the correct tree underlying the distribution increases as edges become weaker, i.e., as $\alpha \rightarrow 0$, because weak edges are harder to detect. As the maximum edge strength β increases, there is a similar increase in sample complexity (as shown by [25, 29] for general bounded degree models). In the context of tree-structured Ising models we give the following theorem (here and throughout, α and β are the bounds $\alpha \leq |\theta_{ij}| \leq \beta$):

Theorem 3.2 (Necessary samples for structure learning). *Given $n < \frac{1}{16} e^{2\beta} \alpha^{-2} \log p$ samples, the worst-case probability of error over trees $\mathsf{T} \in \mathcal{T}$ and distributions*

$P \in \mathcal{P}_{\mathbf{T}}(\alpha, \beta)$ is at least half for any algorithm, i.e.,

$$\inf_{\phi} \sup_{\substack{\mathbf{T} \in \mathcal{T} \\ P \in \mathcal{P}_{\mathbf{T}}(\alpha, \beta)}} P(\phi(X^{1:n}) \neq \mathbf{T}) > 1/2.$$

The proof given in Section 8.1, applies Fano's inequality to a large set of trees that are difficult to distinguish. The next theorem gives an essentially matching sufficient condition.

Theorem 3.3 (Sufficient samples for structure learning). *Fix an arbitrary tree \mathbf{T} and Ising distribution $P \in \mathcal{P}_{\mathbf{T}}(\alpha, \beta)$. If the number of samples is $n > Ce^{2\beta} \max\{\alpha^{-2}, 1\} \log(p/\delta)$ samples, then with probability at least $1 - \delta$ the Chow-Liu algorithm recovers the true tree, i.e., $\mathbf{T}^{\text{CL}} = \mathbf{T}$.*

The proof is presented in Section 8.2. Note that Theorems 3.2 and 3.3 give tight bounds (up to numerical constant) for the sample complexity of learning the tree structure for Ising models with zero external field. The necessary number of samples increases as the minimum edge weight α decreases. Hence, if edges can be arbitrarily weak, it is impossible to learn the tree given a bounded number of samples. If the goal is merely to make accurate predictions, it is natural to seek a less stringent, approximate notion of learning.

Several papers consider learning a model that is close in KL-divergence, e.g. [1, 18, 28]. This corresponds to the log-likelihood loss. The sample complexity for learning a model to within constant KL-divergence ϵ scales at least *linearly* in the number of variables p , an unrealistic requirement in the high-dimensional setting of interest. Using number of samples scaling logarithmically in dimension requires relaxing the KL-divergence to scale as ϵp , but this does not imply a non-trivial guarantee on the quality of approximation for marginals of few variables (as done in this paper).

In the next section we study estimation with respect to the small-set TV loss function, which captures prediction accuracy of the model. We will see that the associated sample complexity is independent of the minimum coupling strength α of the original model.

3.2. Learning for inference

For a subset of nodes $\mathcal{S} \subseteq [p]$ we consider the marginal distribution over variables $X_{\mathcal{S}}$, which can be written as $P_{\mathcal{S}}(x_{\mathcal{S}}) = \sum_{x_{V \setminus \mathcal{S}}} P(x)$. Given two distributions P, Q on the same space, for each $k \geq 1$ the maximum total variation distance over all size k marginals is denoted by $\mathcal{L}^{(k)}(P, Q) \triangleq \sup_{\mathcal{S}: |\mathcal{S}|=k} d_{\text{TV}}(P_{\mathcal{S}}, Q_{\mathcal{S}})$. Note that $\mathcal{L}^{(k)}$ is non-decreasing in k . One can check that $\mathcal{L}^{(k)}$ satisfies the triangle inequality: for any three distributions P, Q, R ,

$$\mathcal{L}^{(k)}(P, Q) + \mathcal{L}^{(k)}(Q, R) \geq \mathcal{L}^{(k)}(P, R). \quad (3.3)$$

Closeness of P and Q in $\mathcal{L}^{(k)}$ implies that the respective posteriors conditioned on subsets of variables of size $k - 1$ are close on average. To see this, suppose that we wish to compute $P(X_i = + | X_{\mathcal{S}})$. We measure performance of the

approximation Q by the expected magnitude of error $|P(x_i = +|X_S) - Q(x_i = +|X_S)|$ averaged over X_S :

$$\begin{aligned} \mathbb{E}_{X_S} |P(X_i = +|X_S) - Q(X_i = +|X_S)| \\ &= \sum_{x_S} |P(X_i = +, X_S = x_S) - Q(X_i = +|X_S = x_S)P(X_S = x_S)| \\ &\leq \sum_{x_S} |P(X_i = +, X_S = x_S) - Q(X_i = +, X_S = x_S)| + \sum_{x_S} |Q(x_S) - P(x_S)| \\ &\leq 2\mathcal{L}^{(|S|+1)}(P, Q). \end{aligned}$$

where the last inequality is a consequence of monotonicity of $\mathcal{L}^{(k)}$ in k .

In this paper we focus on $\mathcal{L}^{(2)}$. Note that for the trees $\mathsf{T}, \tilde{\mathsf{T}} \in \mathcal{T}$ and distributions $P \in \mathcal{P}_{\mathsf{T}}$ and $\tilde{P} \in \mathcal{P}_{\tilde{\mathsf{T}}}$, we have

$$\begin{aligned} \mathcal{L}^{(2)}(P, \tilde{P}) &= \sup_{w, \tilde{w} \in \mathcal{V}} \frac{1}{2} \sum_{x_w, x_{\tilde{w}} \in \{-, +\}^2} |P(x_w, x_{\tilde{w}}) - \tilde{P}(x_w, x_{\tilde{w}})| \\ &= \sup_{w, \tilde{w} \in \mathcal{V}} \frac{1}{2} \left| \prod_{e \in \text{path}_{\mathsf{T}}(w, \tilde{w})} \mu_e - \prod_{e' \in \text{path}_{\tilde{\mathsf{T}}}(w, \tilde{w})} \tilde{\mu}_{e'} \right|, \end{aligned} \quad (3.4)$$

where for $e = (i, j) \in \mathcal{E}_{\mathsf{T}}$, $\mu_e = \mathbb{E}_P X_i X_j$ and for $e' = (i, j) \in \mathcal{E}_{\tilde{\mathsf{T}}}$, $\tilde{\mu}_{e'} = \mathbb{E}_{\tilde{P}} X_i X_j$. Distribution $P \in \mathcal{P}_{\mathsf{T}}$ has singleton uniform marginals (i.e., $P(x_i = +) = 1/2$) and so $P(x_w, x_{\tilde{w}}) = [1 + x_w x_{\tilde{w}} \mathbb{E}_P[X_w X_{\tilde{w}}]]/4$. Also, from the Ising definition (1.1) one can check that $\mathbb{E}_P[X_w X_{\tilde{w}}] = \prod_{e \in \text{path}_{\mathsf{T}}(w, \tilde{w})} \mu_e$. The same holds for $\tilde{P} \in \mathcal{P}_{\tilde{\mathsf{T}}}$, which gives (3.4).

As discussed in detail in section 3.3, bounding $\mathcal{L}^{(2)}$ when $\mathsf{T} \neq \tilde{\mathsf{T}}$ is challenging since these products of correlations are in general computed along different paths as imposed by Equation (3.4).

3.3. Main result

Our main contribution is to provide upper and lower bounds on the number of samples required to get accurate pairwise marginals. To get the upper bound on the number of samples, we study the Chow-Liu algorithm and bound the expression (3.4). The Chow-Liu algorithm recovers the maximum likelihood tree minimizing the expected zero-one loss in (3.2), but as shown in Theorem 3.4, the maximum likelihood tree also performs well in terms of accuracy of pairwise marginals.

Recall that $\Pi_{\mathsf{T}}(P) = \arg \min_{Q \in \mathcal{P}(\mathsf{T})} D(P \| Q)$ is the reverse information projection of the distribution P onto the zero-field Ising models on tree T .

Theorem 3.4 (Learning for inference using Chow-Liu algorithm). *For $\mathsf{T} \in \mathcal{T}$, let the distribution $P \in \mathcal{P}_{\mathsf{T}}(0, \beta)$. Given $n > C \max\{e^{2\beta} \log \frac{p}{\delta}, \eta^{-2} \log \frac{p}{\eta\delta}\}$ samples, if the tree T^{CL} is the Chow-Liu tree as defined in (3.1), then with probability at least $1 - \delta$ we have $\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\text{CL}}}(\hat{P})) < \eta$.*

The main challenge is that the number of samples available in this regime is not sufficient for structure learning, as can be seen by comparing with Theorem 3.2. This means that accurate inference must be performed *using possibly the wrong tree*. To guarantee accurate pairwise marginals using the wrong tree, we characterize the combinatorial and statistical properties of the mistakes made by the algorithm in detecting edges. A sketch of the proof is provided in Section 5.

We also provide a lower bound on the necessary number of samples to guarantee small $\mathcal{L}^{(2)}$ loss.

Theorem 3.5 (Samples necessary for small ssTV). *Suppose one observes $n < \eta^{-1} \operatorname{atanh}^{-1}(\eta) \log p$ samples. Then the worst-case probability of $\mathcal{L}^{(2)}$ loss greater than η taken over trees $T \in \mathcal{T}$ and distributions $P \in \mathcal{P}_T(\alpha, \beta)$ is at least half for any algorithm, i.e.,*

$$\inf_{\phi} \sup_{\substack{T \in \mathcal{T} \\ P \in \mathcal{P}_T(\alpha, \beta)}} \mathbb{P} \left[\mathcal{L}^{(2)}(P, \phi(X^{1:n})) > \eta \right] > 1/2.$$

The proof is provided in Section 8.2.

4. Illustrative example: three node Markov chain

A Markov chain with three nodes captures a few of the key ideas developed in this paper. Let $p = 3$ and $P(X_1, X_2, X_3)$ be represented by the tree T_1 in Figure 1 in which $X_1 \leftrightarrow X_2 \leftrightarrow X_3$ form a Markov chain with correlations μ_{12} and μ_{23} . Without loss of generality, we assume $\mu_{12}, \mu_{23} > 0$. Suppose that for some small value ϵ , $\mu_{12} = 1 - \mu_{23} = \epsilon$.

Given n samples from P , the empirical correlations $\hat{\mu}_{12}$, $\hat{\mu}_{23}$ and $\hat{\mu}_{13}$ between each pair of random variables are concentrated around $\mu_{12} = \epsilon$, $\mu_{23} = 1 - \epsilon$ and $\mu_{13} = \mu_{12}\mu_{23} = \epsilon(1 - \epsilon)$. Let $\hat{\mu}_{12} = \mu_{12} + z_{12}$, $\hat{\mu}_{23} = \mu_{23} + z_{23}$ and $\hat{\mu}_{13} = \mu_{12}\mu_{23} + z_{13}$, where the fluctuations of z_{12} , z_{23} and z_{13} shrink as n grows. It is useful to imagine the typical fluctuations of z_{ij} to be on the order $\epsilon/10$.

Since $\max\{\mu_{12}, \mu_{13}\} = \max\{\epsilon, \epsilon(1 - \epsilon)\} = \epsilon \ll \mu_{23}$, concentration bounds guarantee that with high probability $\hat{\mu}_{23} > \max\{\hat{\mu}_{12}, \hat{\mu}_{13}\}$ and the (greedy implementation of) Chow-Liu algorithm described in (3.1) adds the edge (2, 3) to T^{CL} . But since $\mu_{12} - \mu_{13} = \epsilon^2$ is smaller than fluctuations of z_{12} and z_{13} there is no guarantee that $\hat{\mu}_{12} > \hat{\mu}_{13}$. In particular, if $z_{13} - z_{12} > \epsilon^2$, then $T^{\text{CL}} = T_3$.

The preceding discussion provides the intuition underlying a statistical characterization of the possible errors made by the Chow-Liu algorithm. Later on in the proof, we provide a lower bound $\tau(n, \beta, \delta)$ so that using n samples, any edge e with $|\mu_e| \geq \tau$ is recovered by the Chow-Liu algorithm with probability at least $1 - \delta$. Conversely, if there is a mistake made by Chow-Liu algorithm such that $e \in T$ but $e \notin T^{\text{CL}}$, then $|\mu_e| \leq \tau$. This lower bound is going to play a key role in bounding the ssTV $\mathcal{L}^{(2)}$ for T^{CL} whether or not it is equal to T .

In the regime where n is not large enough to guarantee the correct recovery of all the edges in the tree, there are two natural strategies:

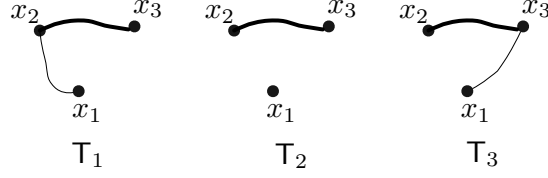


FIG 1. The original distribution is described by T_1 . The width of the edges indicate the coupling strength between the nodes. The truncation algorithm would recover $\hat{T} = T_2$ as the correlation between node x_1 and all the other nodes is not strong enough to recover any edge confidently. Running Chow-Liu algorithm would recover T^{CL} to be T_1 or T_2 depending on the realization of the samples.

- I. *Truncation Algorithm:* The algorithm finds the maximum weight spanning forest of the complete graph, where the weight of each edge e is $|\hat{\mu}_e| - \tau$, for a specified threshold τ . This process is equivalent to finding the maximum weight spanning tree and then truncating the edges with weight smaller than τ . The purpose of this algorithm is recovering a forest F which is a good approximation of the original tree T in the sense that $\mathcal{E}_F \subseteq \mathcal{E}_T$. The edges missing from the forest are all weak edges in the original tree that cannot be recovered reliably using n samples. The appropriate value of τ can be determined based on concentrations inequalities. The details of this algorithm and its sample complexity will be discussed in Section 6.
- II. *Chow-Liu Algorithm:* We can use the Chow-Liu tree as our estimated structure despite the fact that it may well return the incorrect tree.

Under the assumptions given for our three-node example, the truncation algorithm would return $\hat{T} = T_2$ in Figure 1, whereas Chow-Liu would give $\hat{T} = T_1$ or $\hat{T} = T_3$. We compare the *graph estimation error*, defined as $\mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(P))$, for the above cases:

$$\begin{aligned}
 \hat{T} = T_1 & \rightarrow \mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(P)) = 0 \\
 \hat{T} = T_2 & \rightarrow \mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(P)) = |\mu_{12}| = \epsilon \\
 \hat{T} = T_3 & \rightarrow \mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(P)) = |\mu_{12}|(1 - \mu_{23}^2) = \epsilon^2(2 - \epsilon)
 \end{aligned}$$

Evidently, the loss due to graph estimation error in the truncation algorithm is bigger than Chow-Liu algorithm whether or not the latter recovers the true tree. This is because the Chow-Liu algorithm does not make arbitrary errors in estimating the tree: errors happen when both the original tree and the estimated tree describe the original distribution rather well. The formal representation of this statement is the main result in this paper and is given in Theorem 3.4.

5. Outline of Proof

We now sketch the argument for the main result, Theorem 3.4, guaranteeing accurate pairwise marginals in the Chow-Liu tree. The starting point is an application of the triangle inequality (3.3):

$$\mathcal{L}^{(2)}(P, \Pi_{\text{Tcl}}(\hat{P})) \leq \mathcal{L}^{(2)}(P, \Pi_{\text{Tcl}}(P)) + \mathcal{L}^{(2)}(\Pi_{\text{Tcl}}(P), \Pi_{\text{Tcl}}(\hat{P})). \quad (5.1)$$

The first term on the right-hand side of Equation (5.1) addresses the error in estimation of pairwise correlations due to the difference in the structure of \mathbf{T} and \mathbf{T}^{cl} . Looking at Equation (3.4), for each pair of nodes $u, v \in \mathcal{V}$, $\text{path}_{\mathbf{T}}(u, v)$ and $\text{path}_{\text{Tcl}}(u, v)$ must be compared. We call this first term the *loss due to graph estimation error*.

The second term on the right-hand side of Equation (3.4) addresses the propagation of error due to the parameter estimation error. Recall that the estimated parameters on the Chow-Liu tree are derived by matching correlations to the empirical values.

Theorem 3.4 follows by separately bounding each term on the right-hand side of Equation (5.1).

Proposition 5.1 (Loss due to edge estimation error). *Given*

$$n > C \max\{e^{2\beta} \log \frac{p}{\delta}, \eta^{-2} \log \frac{p}{\eta\delta}\}$$

samples, with probability at least $1 - \delta$ we have $\mathcal{L}^{(2)}(\Pi_{\text{Tcl}}(P), \Pi_{\text{Tcl}}(\hat{P})) < \eta$.

Proposition 5.2 (Loss due to graph estimation error). *Given*

$$n > C \max\{e^{2\beta}, \eta^{-2}\} \log \frac{p}{\delta}$$

samples, with probability at least $1 - \delta$ we have $\mathcal{L}^{(2)}(P, \Pi_{\text{Tcl}}(P)) < \eta$.

These two propositions are proved in full detail in Sections 7.1 and 7.2. In the remainder of this section we define probabilistic events of interest and sketch the proofs. Proposition 5.1 has a relatively straightforward proof, while the major technical contribution is in Proposition 5.2. The latter uses a careful inductive argument to control the errors arising from recovery of an incorrect tree.

We define three highly probable events $\mathbf{E}^{\text{corr}}(\epsilon)$, $\mathbf{E}^{\text{strong}}(\epsilon)$ and $\mathbf{E}^{\text{cascade}}(\epsilon)$ as follows. Let $\mathbf{E}^{\text{corr}}(\epsilon)$ be the event that all empirical correlations within ϵ of population values:

$$\mathbf{E}^{\text{corr}}(\epsilon) = \left\{ \sup_{w, \tilde{w} \in \mathcal{V}} |\mu_{w, \tilde{w}} - \hat{\mu}_{w, \tilde{w}}| \leq \epsilon \right\}. \quad (5.2)$$

Let $\tau = 4\epsilon/\sqrt{1 - \tanh \beta}$ and

$$\mathcal{E}_{\mathbf{T}}^{\text{strong}} = \{(i, j) \in \mathcal{E}_{\mathbf{T}} : |\tanh(\theta_{i,j})| \geq \tau\} \quad (5.3)$$

consist of the set of strong edges in tree T and let T^{CL} be the Chow-Liu tree as in Equation (3.1). Let $\mathsf{E}^{\text{strong}}(\epsilon)$ be the event that all strong edges in T as defined in (5.3) are recovered using the Chow-Liu algorithm:

$$\mathsf{E}^{\text{strong}}(\epsilon) = \{ \mathcal{E}_{\mathsf{T}}^{\text{strong}} \subset \mathcal{E}_{\mathsf{T}^{\text{CL}}} \}. \quad (5.4)$$

Finally, define the event

$$\mathsf{E}^{\text{cascade}}(\epsilon) = \{ \mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}}(\hat{P})) \leq \epsilon \}. \quad (5.5)$$

Recall that P factorizes according to T and \hat{P} is the empirical distribution which is not factorized according to any tree. This event controls *cascades* of errors in correlations computed along paths as per (3.4).

Lemma 5.3. *The following hold:*

$$\begin{aligned} \Pr[\mathsf{E}^{\text{corr}}(\epsilon)] &\geq 1 - 2p^2 \exp(-n\epsilon^2/2) \\ \Pr[\mathsf{E}^{\text{strong}}(\epsilon)] &\geq 1 - 2p^2 \exp(-n\epsilon^2/2) \\ \Pr[\mathsf{E}^{\text{cascade}}(\epsilon)] &\geq 1 - 4p^2/\epsilon \exp(-\epsilon^2 n/32). \end{aligned}$$

This Lemma is the restatement of Lemmas 9.2, 9.9 and 9.12.

Sketch of proof of Proposition 5.1. The proof of Proposition 5.1 entails showing that on the event $\mathsf{E}^{\text{corr}}(\epsilon) \cap \mathsf{E}^{\text{strong}}(\epsilon) \cap \mathsf{E}^{\text{cascade}}(\gamma)$ with $\epsilon < \min\{e^{-\beta}/20, \eta/3\}$ and $\gamma \leq \eta/3$ we have the desired inequality $\mathcal{L}^{(2)}(\Pi_{\mathsf{T}^{\text{CL}}}(P), \Pi_{\mathsf{T}^{\text{CL}}}(\hat{P})) < \eta$. According to Lemma 5.3 this happens with probability at least $1 - 3\delta$ with $n > \max\{800e^{2\beta} \log \frac{4p^2}{\delta}, 300\eta^{-2} \log \frac{12p^2}{\eta\delta}\}$ samples, which gives the proposition.

To bound $\mathcal{L}^{(2)}(\Pi_{\mathsf{T}^{\text{CL}}}(P), \Pi_{\mathsf{T}^{\text{CL}}}(\hat{P}))$, we consider the net error due to parameter estimation errors in every path in T^{CL} . On the event $\mathsf{E}^{\text{cascade}}(\gamma)$, according to (5.5), the end-to-end error for each path in $\mathcal{E}_{\mathsf{T}} \cap \mathcal{E}_{\mathsf{T}^{\text{CL}}}$ is bounded by γ .

Next, we study the end-to-end error due to parameter estimation in paths containing (falsely added) edges in $\mathcal{E}_{\mathsf{T}^{\text{CL}}} \setminus \mathcal{E}_{\mathsf{T}}$. For any pair of nodes w, \tilde{w} denote by t the number of falsely added edges $|\text{path}_{\mathsf{T}^{\text{CL}}}(w, \tilde{w}) \setminus \mathcal{E}_{\mathsf{T}}|$. As discussed in the proof, these edges correspond to missed edges in T and thus $\mathsf{E}^{\text{strong}}(\epsilon)$ guarantees that these edges are weak (as defined in (5.3)). These t weak edges break up $\text{path}_{\mathsf{T}^{\text{CL}}}(w, \tilde{w})$ into at most $t+1$ contiguous segments $\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_t$ each entirely within $\mathcal{E}_{\mathsf{T}} \cap \mathcal{E}_{\mathsf{T}^{\text{CL}}}$.

The error on each segment $\mathcal{F}_i, 0 \leq i \leq t$, is bounded by γ as already discussed, but now there are $t+1$ such segments and the errors add up. This effect is counterbalanced by the fact that the falsely added edges are weak and hence scale down the error in a multiplicative fashion. The error $\left| \prod_{e \in \text{path}_{\mathsf{T}^{\text{CL}}}(w, \tilde{w})} \hat{\mu}_e - \prod_{e \in \text{path}_{\mathsf{T}^{\text{CL}}}(w, \tilde{w})} \mu_e \right|$ dependence on t therefore has a linearly increasing factor and an exponentially decaying factor. If $\epsilon \leq e^{-\beta}/20$, then this gives a uniform upper bound of $3 \max\{\epsilon, \gamma\}$ for this quantity for any number t of falsely added edges. Putting it all together shows that if $\epsilon \leq \min\{e^{-\beta}/20, \eta/3\}$ and $\gamma \leq \eta/3$ then $\mathcal{L}^{(2)}(\Pi_{\mathsf{T}^{\text{CL}}}(P), \Pi_{\mathsf{T}^{\text{CL}}}(\hat{P})) < \eta$ on the events $\mathsf{E}^{\text{corr}}(\epsilon) \cap \mathsf{E}^{\text{strong}}(\epsilon) \cap \mathsf{E}^{\text{cascade}}(\gamma)$.

Sketch of proof of Proposition 5.2. The goal is to show that the event $\mathbf{E}^{\text{corr}}(\epsilon) \cap \mathbf{E}^{\text{strong}}(\epsilon)$ with $\epsilon < \min\{\frac{\eta}{16}, e^{-\beta}/24\}$ implies $\mathcal{L}^{(2)}(P, \Pi_{\mathcal{T}^{\text{cl}}}(P)) < \eta$. Lemma 5.3 then implies that $n > 1200 \max\{e^{2\beta}, \eta^{-2}\} \log \frac{2p^2}{\delta}$ suffices for $\mathbf{E}^{\text{corr}}(\epsilon) \cap \mathbf{E}^{\text{strong}}(\epsilon)$ to occur with probability at least $1 - 2\delta$.

The proof of the proposition sets up a careful induction on the distance between nodes for which we wish to bound the error in correlation. The key technical ingredient is a combinatorial lemma relating trees \mathcal{T}^{cl} and \mathcal{T} . Lemma 10.1 studies any two spanning trees on p nodes and the paths between two given nodes w and \tilde{w} . It proves the existence of at least one pair of corresponding edges $f \in \text{path}_{\mathcal{T}}(w, \tilde{w})$ and $g \in \text{path}_{\mathcal{T}^{\text{cl}}}(w, \tilde{w})$ satisfying a collection of properties illustrated in Figure 7.2 (and specified in the Lemma). A consequence is that the true correlation across g according to P can be expressed as a function of the correlation on f as $\mu_g = \mu_f \mu_A \mu_C \mu_{\tilde{A}} \mu_{\tilde{C}}$, hence $|\mu_g| \leq |\mu_f|$. But $|\hat{\mu}_f| \geq |\hat{\mu}_g|$, since the Chow-Liu algorithm chose g in \mathcal{T}^{cl} instead of f .

In the proof of Proposition 5.2 a recurrence $\Delta(d)$ is defined. $\Delta(d)$ provides an upper bound for the error due to graph estimation error for any pair of nodes w, \tilde{w} with $|\text{path}_{\mathcal{T}^{\text{cl}}}(w, \tilde{w})| = d$. Using the relationship between the edges f and g , the paths $\text{path}_{\mathcal{T}}(w, \tilde{w})$ and $\text{path}_{\mathcal{T}^{\text{cl}}}(w, \tilde{w})$ are compared. It is proved that the recurrence $\Delta(d)$ increases linearly and decays exponentially with d . With $\epsilon < \min\{\frac{\eta}{16}, e^{-\beta}/24\}$, convergence of $\Delta(d)$ with the upper bound η is shown.

6. Truncation Algorithm

In the regime where exact recovery of the tree is impossible, a reasonable goal is to instead find a forest approximation to the tree. In this section we analyze a natural truncation algorithm, which thresholds to zero edges with correlation below a specified value τ .

There is extensive literature on estimating forests in the fully observed setting of this paper, including [28, 18]. [21] studied the problem of learning phylogenetic forests, where samples are only observed at the leaves of the tree. They quantified the idea that most edges of phylogenies are easy to reconstruct, and in the regime that the sample complexity of the structure learning is too high, they instead reconstruct several edge disjoint trees. An upper bound on the number of edges necessary to glue together the forest to get the original tree is provided as a function of number of leaves, the minimum edge weight, and the metric distortion bounds. Our results in this section are consistent with the asymptotic conditions on the thresholds given in [28] for the general distributions over trees.

The truncation algorithm considered in this section thresholds the edges with correlations below $\tau = \frac{4\epsilon}{\sqrt{1-\tanh\beta}}$ for $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$ to zero. This is equivalent to finding the maximum-weight spanning forest over the complete graph with edge weights $|\hat{\mu}_e| - \tau - \epsilon$. The output $\hat{\mathcal{T}} = (\mathcal{V}, \mathcal{E}_{\hat{\mathcal{T}}})$ is a truncated version of \mathcal{T}^{cl} such that the empirical correlation between any pair of nodes $(i, j) \in \mathcal{E}_{\hat{\mathcal{T}}}$ satisfies $|\hat{\mu}_{ij}| \geq \tau + \epsilon$.

Proposition 6.1. *Given $n > C \max\{e^{2\beta}/\eta^2 \log \frac{p}{\delta}, 1/\eta^2 \log \frac{p}{\eta\delta}\}$ samples, the*

truncation algorithm guarantees that with probability $1 - \delta$ we have $\mathcal{E}_{\hat{T}} \subseteq \mathcal{E}_T$ and $\mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(\hat{P})) < \eta$.

Proof. We will provide ϵ and γ such that on the event $\mathbf{E}^{\text{corr}}(\epsilon) \cap \mathbf{E}^{\text{strong}}(\epsilon) \cap \mathbf{E}^{\text{cascade}}(\gamma)$ the ssTV bound holds. All the weak edges $e \in \mathcal{E}_T$ such that $|\mu_e| \leq \tau$ (for which we cannot guarantee correct recovery by Chow-Liu algorithm) have $|\hat{\mu}_e| \leq \tau + \epsilon$ on event $\mathbf{E}^{\text{corr}}(\epsilon)$, hence these edges are truncated by the thresholding process. It follows that $\mathcal{E}_{\hat{T}} \subseteq \mathcal{E}_T$.

On event $\mathbf{E}^{\text{strong}}(\epsilon)$, the strong edges defined in Equation (5.3) are recovered by the Chow-Liu algorithm, i.e., for all $e \in \mathcal{E}_T$, with $|\mu_e| \geq \tau + 2\epsilon$, $e \in \mathcal{T}^{\text{CL}}$. On event $\mathbf{E}^{\text{corr}}(\epsilon)$, these edges also satisfy $|\hat{\mu}_e| \geq \tau + \epsilon$ and these edges are retained. This implies that all of the strong edges in the original tree T (with $|\mu_e| \geq \tau + 2\epsilon$) are recovered correctly by the truncation algorithm (i.e., $\mathcal{E}_T^{\text{strong}} \subseteq \mathcal{E}_{\hat{T}}$).

To bound the ssTV, we decompose into two terms as per (5.1). We first study the loss due to graph estimation error $\mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(P))$. For any pair of nodes w, \tilde{w} in \hat{T} , since $\mathcal{E}_{\hat{T}} \subseteq \mathcal{E}_T$, if there is a path between them, then this is the same as the path between them in T . Hence looking at (3.4), $\mathbb{E}_P[X_w X_{\tilde{w}}] = \mathbb{E}_{\Pi_{\hat{T}}(P)}[X_w X_{\tilde{w}}]$. If there is no path between w, \tilde{w} in \hat{T} , then $\mathbb{E}_{\Pi_{\hat{T}}(P)}[X_w X_{\tilde{w}}] = 0$. This implies that there is an edge $e \in \text{path}_T(w, \tilde{w})$ such that $e \notin \mathcal{E}_{\hat{T}}$. As proved, all the missing edges in \hat{T} are weak edges, hence $|\mu_e| \leq \tau + 2\epsilon$. This implies that $|\mathbb{E}_P[X_w X_{\tilde{w}}]| \leq |\mu_e| \leq \tau + 2\epsilon$ and it follows that $\mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(P)) \leq \tau + 2\epsilon$.

To bound $\mathcal{L}^{(2)}(\Pi_{\hat{T}}(P), \Pi_{\hat{T}}(\hat{P}))$, we again consider an arbitrary pair of nodes w, \tilde{w} with a path between them in \hat{T} . Since $\mathcal{E}_{\hat{T}} \subseteq \mathcal{E}_T$, under the event $\mathbf{E}^{\text{cascade}}(\gamma)$, $|\mathbb{E}_{\Pi_{\hat{T}}(P)}[X_w X_{\tilde{w}}] - \mathbb{E}_{\Pi_{\hat{T}}(\hat{P})}[X_w X_{\tilde{w}}]| \leq \gamma$. If there is no path between w, \tilde{w} in \hat{T} , then $\mathbb{E}_{\Pi_{\hat{T}}(P)}[X_w X_{\tilde{w}}] = \mathbb{E}_{\Pi_{\hat{T}}(\hat{P})}[X_w X_{\tilde{w}}] = 0$. Hence, $\mathcal{L}^{(2)}(\Pi_{\hat{T}}(P), \Pi_{\hat{T}}(\hat{P})) \leq \gamma$.

Combining the two error terms gives $\mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(\hat{P})) \leq \tau + 2\epsilon + \gamma$. If $\epsilon \leq \eta e^{-\beta}/8$ and $\gamma \leq \eta/4$ then $\tau = 4\epsilon/\sqrt{1 - \tanh \beta} \leq 4\epsilon e^{\beta} \leq \eta/8$ and $\mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(\hat{P})) \leq \tau + 2\epsilon + \gamma \leq \eta$. By Lemma 5.3, using $n > \max\{128e^{2\beta}/\eta^2 \log \frac{2p^2}{\delta}, 512/\eta^2 \log \frac{32p^2}{\eta\delta}\}$ samples, the event $\mathbf{E}^{\text{corr}}(\epsilon) \cap \mathbf{E}^{\text{strong}}(\epsilon) \cap \mathbf{E}^{\text{cascade}}(\gamma)$ happens with probability $1 - 3\delta$ with $\epsilon \leq \eta e^{-\beta}/8 \leq \eta/8$ and $\gamma < \eta/4$ which proves the result. \square

Comparing the above statement with Proposition 5.2 shows that in general the loss due to graph estimation using the truncation algorithm is (asymptotically) strictly greater than the loss incurred by the Chow-Liu algorithm.

7. Proof of main result

As stated in Section 5, Theorem 3.3 is a direct consequence of Propositions 5.1 and 5.2, proved in Sections 7.1 and 7.2.

7.1. Proof of Proposition 5.1

By Lemma 5.3, the event $\mathbf{E}(\epsilon, \gamma) := \mathbf{E}^{\text{corr}}(\epsilon) \cap \mathbf{E}^{\text{strong}}(\epsilon) \cap \mathbf{E}^{\text{cascade}}(\gamma)$ occurs with probability at least $1 - 3\delta$ provided $n > \max\{2/\epsilon^2 \log(4p^2/\delta), 32/\gamma^2 \log(4p^2/\gamma\delta)\}$.

By assumption $n > \max\{800e^{2\beta} \log \frac{4p^2}{\delta}, 300\eta^{-2} \log \frac{12p^2}{\eta\delta}\}$, hence $\gamma \leq \eta/3$ and $\epsilon \leq \min\{e^{-\beta}/20, \eta/3\}$. We also assume $\epsilon < \gamma$ for given n number of samples and δ . It remains to show that on the event $\mathbf{E}(\epsilon, \gamma)$ if $\epsilon \leq \min\{e^{-\beta}/20, \eta/3\}$ and $\gamma \leq \eta/3$, then $\mathcal{L}^{(2)}(\Pi_{\text{TCL}}(P), \Pi_{\text{TCL}}(\hat{P})) < \eta$.

Let $\tau = 4\epsilon/\sqrt{1 - \tanh \beta}$ be the threshold used in (5.3) to define $\mathcal{E}_{\text{T}}^{\text{strong}}$. For any pair of nodes w, \tilde{w} , consider $\text{path}_{\text{TCL}}(w, \tilde{w})$. Let $0 \leq t < p$ be the number of weak edges $e_1, \dots, e_t \in \text{path}_{\text{TCL}}(w, \tilde{w})$ such that $|\mu_{e_i}| < \tau$. There are at most $t+1$ contiguous sub-paths in $\text{path}_{\text{TCL}}(w, \tilde{w})$ consisting of strong edges. We call these segments $\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_t$ where $\mathcal{F}_i \subseteq \text{path}_{\text{TCL}}(w, \tilde{w})$ could be empty in which case we define $\mu_{\mathcal{F}_i} = \hat{\mu}_{\mathcal{F}_i} = 1$. By definition, all edges $f \in \mathcal{F}_i$ are strong, i.e., $|\mu_f| \geq \tau$.

According to Lemma 9.9, under the event $\mathbf{E}^{\text{strong}}(\epsilon)$ all strong edges as in (5.3) are recovered in TCL . Thus, $\mathcal{F}_i \subseteq \mathcal{E}_{\text{T}}$, which guarantees $|\hat{\mu}_{\mathcal{F}_i} - \mu_{\mathcal{F}_i}| \leq \gamma$ under the event $\mathbf{E}^{\text{cascade}}(\gamma)$.

Note that Lemma 9.12 gives the bound for $t = 0$ which corresponds to no weak edges in the path in TCL . For $t \geq 1$ we have:

$$\begin{aligned}
& \left| \prod_{e \in \text{path}_{\text{TCL}}(w, \tilde{w})} \hat{\mu}_e - \prod_{e \in \text{path}_{\text{TCL}}(w, \tilde{w})} \mu_e \right| \\
& \stackrel{(a)}{=} \left| \hat{\mu}_{\mathcal{F}_0} \prod_{i=1}^t \hat{\mu}_{\mathcal{F}_i} \hat{\mu}_{e_i} - \mu_{\mathcal{F}_0} \prod_{i=1}^t \mu_{\mathcal{F}_i} \mu_{e_i} \right| \\
& \stackrel{(b)}{\leq} |\hat{\mu}_{\mathcal{F}_0} - \mu_{\mathcal{F}_0}| \prod_{j=1}^t |\mu_{\mathcal{F}_j} \mu_{e_j}| \\
& \quad + \sum_{i=1}^t |\hat{\mu}_{\mathcal{F}_i} \hat{\mu}_{e_i} - \mu_{\mathcal{F}_i} \mu_{e_i}| \cdot |\hat{\mu}_{\mathcal{F}_0}| \prod_{j=1}^{i-1} |\hat{\mu}_{\mathcal{F}_j} \hat{\mu}_{e_j}| \prod_{j=i+1}^t |\mu_{\mathcal{F}_j} \mu_{e_j}| \\
& \stackrel{(c)}{\leq} \gamma \tau^t + (\tau + \epsilon)^{t-1} \sum_{i=1}^t |\hat{\mu}_{\mathcal{F}_i} \hat{\mu}_{e_i} - \mu_{\mathcal{F}_i} \mu_{e_i}| \\
& \stackrel{(d)}{\leq} \gamma \tau^t + (\tau + \epsilon)^{t-1} \left[\sum_{i=1}^t |\mu_{\mathcal{F}_i} (\hat{\mu}_{e_i} - \mu_{e_i})| + |\hat{\mu}_{e_i} (\hat{\mu}_{\mathcal{F}_i} - \mu_{\mathcal{F}_i})| \right] \\
& \stackrel{(e)}{\leq} (\tau + \epsilon)^{t-1} (2t+1) \max\{\gamma, \epsilon\} \stackrel{(f)}{\leq} (\tau + \epsilon)^{t-1} (2t+1) \eta/3 \\
& \stackrel{(g)}{\leq} \left[\frac{4\epsilon}{\sqrt{1 - \tanh \beta}} + \epsilon \right]^{t-1} (2t+1) \eta/3 \stackrel{(h)}{\leq} [4\epsilon e^{\beta} + \epsilon]^{t-1} (2t+1) \gamma \\
& \stackrel{(i)}{\leq} (5\epsilon e^{\beta})^{t-1} (2t+1) \gamma \stackrel{(j)}{\leq} \frac{2t+1}{4^{t-1}} \eta/3 \stackrel{(k)}{\leq} \eta.
\end{aligned}$$

In (a) we use $\text{path}_{\text{TCL}}(w, \tilde{w}) = \{\mathcal{F}_0, e_1, \dots, \mathcal{F}_t, e_t, \mathcal{F}_t\}$. (b) uses the bound $|\prod_{i=1}^n a_i - \prod_{i=1}^n b_i| \leq \sum_{i=1}^n |a_i - b_i| \prod_{j=1}^{i-1} |a_j| \prod_{k=i+1}^n |b_k|$ obtained via telescoping sum and triangle inequality. In (c) we use $|\mu_{\mathcal{F}_i}| \leq 1$, $|\hat{\mu}_{\mathcal{F}_i}| \leq 1$, $|\mu_{e_i}| \leq \tau$, $|\hat{\mu}_{e_i}| \leq \tau + \epsilon$

under $\mathbf{E}^{\text{corr}}(\epsilon)$ and $|\hat{\mu}_{\mathcal{F}_i} - \mu_{\mathcal{F}_i}| \leq \gamma$ under $\mathbf{E}^{\text{cascade}}(\gamma)$. In (d) telescoping sum is used again. In (e), we use $|\hat{\mu}_{\mathcal{F}_i} - \mu_{\mathcal{F}_i}| \leq \gamma$ and $|\hat{\mu}_{e_i} - \mu_{e_i}| \leq \epsilon$. (f) is true under the assumption $\gamma \leq \eta/3$. (g) uses the definition of $\tau = 4\epsilon/\sqrt{1 - \tanh \beta}$. (h) and (i) use inequality $1 - \tanh \beta \geq e^{-2\beta}$ and $\beta \geq 0$. (j) uses the assumption $\epsilon \leq \min\{e^{-\beta}/20, \eta/3\}$ and (k) holds for all $t \geq 1$. \square

7.2. Proof of Proposition 5.2

Recall that P factorizes according to T . The error in correlation between any two variables $X_w, X_{\tilde{w}}$ computed along the path T^{cl} as compared to T is

$$\begin{aligned} \text{error}_{P, \mathsf{T}^{\text{cl}}}(w, \tilde{w}) &= 2 \left| \mathbb{E}_P X_w X_{\tilde{w}} - \mathbb{E}_{\Pi_{\mathsf{T}^{\text{cl}}}(P)} X_w X_{\tilde{w}} \right| \\ &= 2 \left| \prod_{e \in \text{path}_{\mathsf{T}}(w, \tilde{w})} \mu_e - \prod_{e \in \text{path}_{\mathsf{T}^{\text{cl}}}(w, \tilde{w})} \mu_e \right|. \end{aligned} \quad (7.1)$$

Our goal is to bound $\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\text{cl}}}(P)) = \sup_{w, \tilde{w} \in \mathcal{V}} \text{error}_{P, \mathsf{T}^{\text{cl}}}(w, \tilde{w})$.

With $n > 1200 \max\{e^{2\beta}, \eta^{-2}\} \log(2p^2/\delta)$, by Lemma 5.3, the event $\mathbf{E}^{\text{corr}}(\epsilon) \cap \mathbf{E}^{\text{strong}}(\epsilon)$ occurs with probability $1 - 2\delta$ with $\epsilon < \min\{\eta/16, e^{-\beta}/24\}$. In the following we prove that for $\epsilon < \min\{\eta/16, e^{-\beta}/24\}$ on the event $\mathbf{E}^{\text{corr}}(\epsilon) \cap \mathbf{E}^{\text{strong}}(\epsilon)$ we have the inequality $\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\text{cl}}}(P)) < \eta$.

We derive a recurrence on the maximum of $\text{error}_{P, \mathsf{T}^{\text{cl}}}(w, \tilde{w})$ in terms of the distance between the nodes w, \tilde{w} in T^{cl} . Define

$$\Delta(k) \triangleq \sup_{\substack{w, \tilde{w} \in \mathcal{V} \\ |\text{path}_{\mathsf{T}^{\text{cl}}}(w, \tilde{w})| = k}} \text{error}_{P, \mathsf{T}^{\text{cl}}}(w, \tilde{w}).$$

For each $d > 1$, we want to bound $\Delta(d)$ in terms of $\Delta(k)$ for $k < d$. For nodes at distance one, $|\text{path}_{\mathsf{T}^{\text{cl}}}(w, \tilde{w})| = 1$ and it follows that $\text{error}_{P, \mathsf{T}^{\text{cl}}}(w, \tilde{w}) = 0$ from the definition of the projected distribution $\Pi_{\mathsf{T}^{\text{cl}}}(P)$ (matching pairwise marginals across edges). Hence, $\Delta(1) \leq \eta$, and we define $\Delta(0) = 0$.

Note that if $\text{path}_{\mathsf{T}^{\text{cl}}}(w, \tilde{w}) = \text{path}_{\mathsf{T}}(w, \tilde{w})$, then $\text{error}_{P, \mathsf{T}^{\text{cl}}}(w, \tilde{w}) = 0$ (again because correlations are matched across edges). Thus, we assume $\text{path}_{\mathsf{T}^{\text{cl}}}(w, \tilde{w}) \neq \text{path}_{\mathsf{T}}(w, \tilde{w})$. Lemma 10.1 shows the existence of a pair of edges $f = (u, \tilde{u})$ and $g = (v, \tilde{v})$ with the following properties:

- $f \in \text{path}_{\mathsf{T}}(w, \tilde{w})$ and $g \in \text{path}_{\mathsf{T}^{\text{cl}}}(w, \tilde{w})$.
- $f \notin \text{path}_{\mathsf{T}^{\text{cl}}}(w, \tilde{w})$ and $g \notin \text{path}_{\mathsf{T}}(w, \tilde{w})$.
- $f \in \text{path}_{\mathsf{T}}(v, \tilde{v})$ and $g \in \text{path}_{\mathsf{T}^{\text{cl}}}(u, \tilde{u})$.
- $u, v \in \text{SubTree}_{\mathsf{T}, f}(w)$ and $\tilde{u}, \tilde{v} \in \text{SubTree}_{\mathsf{T}, f}(\tilde{w})$.

Here $\text{SubTree}_{\mathsf{T}, f}(w) = \{i \in \mathcal{V}; f \notin \text{path}_{\mathsf{T}}(w, i)\}$ is the set of nodes connected to w in T after removing edge f (see Figure 7.2).

We define several sub-paths (see Figure 7.2):

$$\begin{aligned} \mathcal{A} &= \text{path}_{\mathsf{T}}(u, w) \cap \text{path}_{\mathsf{T}}(u, v), & \mathcal{B} &= \text{path}_{\mathsf{T}}(u, w) \setminus \text{path}_{\mathsf{T}}(u, v), \\ \mathcal{C} &= \text{path}_{\mathsf{T}}(u, v) \setminus \text{path}_{\mathsf{T}}(u, w), & \mathcal{D} &= \text{path}_{\mathsf{T}^{\text{cl}}}(w, v). \end{aligned}$$

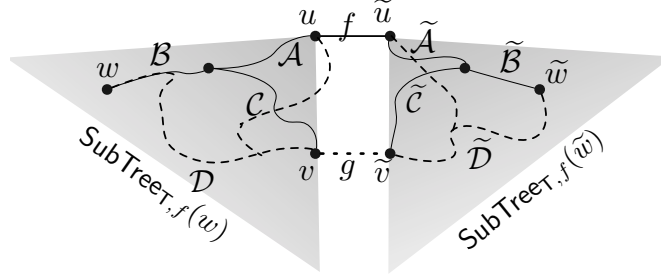


FIG 2. Schematic for the proof of Proposition 5.2. The thin solid lines represent paths in \mathbf{T} and the dashed lines represent paths in \mathbf{T}^{cl} . Edge f is in \mathbf{T} but not in \mathbf{T}^{cl} , while g is in \mathbf{T}^{cl} but not in \mathbf{T} . The sets \mathcal{D} and $\tilde{\mathcal{D}}$ may overlap with $\mathcal{A} \cup \mathcal{B}$ and $\tilde{\mathcal{A}} \cup \tilde{\mathcal{B}}$, respectively. The property that $g \in \text{path}_{\mathbf{T}^{\text{cl}}}(v, \tilde{v})$ is not shown in this figure.

Recall that for set of edges \mathcal{S} , we defined $\mu_{\mathcal{S}} = \prod_{e \in \mathcal{S}} \mu_e$. Since $\text{path}_{\mathbf{T}}(w, v) = \mathcal{B} \cup \mathcal{C}$ and $\mathcal{B} \cap \mathcal{C} = \emptyset$, we have $\mu_{w,v} = \mu_{\mathcal{B}}\mu_{\mathcal{C}}$. Similarly, in $\text{SubTree}_{\mathbf{T},f}(\tilde{w})$ we define

$$\begin{aligned} \tilde{\mathcal{A}} &= \text{path}_{\mathbf{T}}(\tilde{u}, \tilde{w}) \cap \text{path}_{\mathbf{T}}(\tilde{u}, \tilde{v}), & \tilde{\mathcal{B}} &= \text{path}_{\mathbf{T}}(\tilde{u}, \tilde{w}) \setminus \text{path}_{\mathbf{T}}(\tilde{u}, \tilde{v}), \\ \tilde{\mathcal{C}} &= \text{path}_{\mathbf{T}}(\tilde{u}, \tilde{v}) \setminus \text{path}_{\mathbf{T}}(\tilde{u}, \tilde{w}), & \tilde{\mathcal{D}} &= \text{path}_{\mathbf{T}^{\text{cl}}}(\tilde{w}, \tilde{v}). \end{aligned}$$

The sets are defined so that $\text{path}_{\mathbf{T}}(v, \tilde{v}) = \mathcal{C} \cup \mathcal{A} \cup \{f\} \cup \tilde{\mathcal{C}} \cup \tilde{\mathcal{A}}$ where $f = (u, \tilde{u})$ and $g = (v, \tilde{v}) \in \mathcal{E}_{\mathbf{T}^{\text{cl}}}$. Thus, $\mu_g = \mu_f \mu_{\mathcal{A}} \mu_{\mathcal{C}} \mu_{\tilde{\mathcal{A}}} \mu_{\tilde{\mathcal{C}}}$. Since $\text{path}_{\mathbf{T}}(w, \tilde{w}) = \mathcal{A} \cup \mathcal{B} \cup \{f\} \cup \tilde{\mathcal{A}} \cup \tilde{\mathcal{B}}$ and $\text{path}_{\mathbf{T}^{\text{cl}}}(w, \tilde{w}) = \mathcal{D} \cup \{g\} \cup \tilde{\mathcal{D}}$ our goal amounts to finding an upper bound for the quantity $|\mu_{\mathcal{D}} \mu_g \mu_{\tilde{\mathcal{D}}} - \mu_{\mathcal{A}} \mu_{\mathcal{B}} \mu_f \mu_{\tilde{\mathcal{A}}} \mu_{\tilde{\mathcal{B}}}|$.

Lemma 10.3 applied to the edges f and g gives $|\mu_f| - 4\epsilon \leq |\mu_f \mu_{\mathcal{A}} \mu_{\mathcal{C}} \mu_{\tilde{\mathcal{A}}} \mu_{\tilde{\mathcal{C}}}| \leq |\mu_f|$. Thus,

$$|\mu_f| \left(1 - |\mu_{\tilde{\mathcal{C}}}^2 \mu_{\tilde{\mathcal{C}}}^2|\right) \leq 2|\mu_f| (1 - |\mu_{\mathcal{C}} \mu_{\tilde{\mathcal{C}}}|) \leq 2|\mu_f| (1 - |\mu_{\mathcal{A}} \mu_{\mathcal{C}} \mu_{\tilde{\mathcal{A}}} \mu_{\tilde{\mathcal{C}}}|) \leq 8\epsilon. \quad (7.2)$$

Since $f \in \mathcal{E}_{\mathbf{T}}$ and $f \notin \mathcal{E}_{\mathbf{T}^{\text{cl}}}$, under the event $\mathbf{E}^{\text{strong}}(\epsilon)$, f cannot be a strong edge as defined in Equation (5.3). It follows that $|\mu_f| \leq \tau$ for $\tau = 4\epsilon/\sqrt{1 - \tanh \beta}$.

Let $k = |\mathcal{D}| = |\text{path}_{\mathbf{T}^{\text{cl}}}(w, v)|$ and $\tilde{k} = |\tilde{\mathcal{D}}| = |\text{path}_{\mathbf{T}^{\text{cl}}}(\tilde{w}, \tilde{v})|$, so $d = k + \tilde{k} + 1$. By definition of $\Delta(\cdot)$ function:

$$\text{error}_{P, \mathbf{T}^{\text{cl}}}(w, v) = |\mu_{\mathcal{D}} - \mu_{\mathcal{B}} \mu_{\mathcal{C}}| \leq \Delta(k), \quad \text{error}_{P, \mathbf{T}^{\text{cl}}}(\tilde{w}, \tilde{v}) = |\mu_{\tilde{\mathcal{D}}} - \mu_{\tilde{\mathcal{B}}} \mu_{\tilde{\mathcal{C}}}| \leq \Delta(\tilde{k}) \quad (7.3)$$

Using $\mu_g = \mu_A \mu_C \mu_f \mu_{\tilde{A}} \mu_{\tilde{C}}$ and subsequently Equation (7.3),

$$\begin{aligned}
& \text{error}_{P, \text{TCI}}(w, \tilde{w}) \\
&= |\mu_{\mathcal{D}} \mu_g \mu_{\tilde{\mathcal{D}}} - \mu_A \mu_B \mu_f \mu_{\tilde{A}} \mu_{\tilde{B}}| \\
&= |\mu_{\mathcal{D}} \mu_A \mu_C \mu_f \mu_{\tilde{A}} \mu_{\tilde{C}} \mu_{\tilde{\mathcal{D}}} - \mu_A \mu_B \mu_f \mu_{\tilde{A}} \mu_{\tilde{B}}| \\
&= |\mu_A \mu_f \mu_{\tilde{A}}| \cdot |\mu_{\mathcal{D}} \mu_C \mu_{\tilde{C}} \mu_{\tilde{\mathcal{D}}} - \mu_B \mu_{\tilde{B}}| \\
&= |\mu_A \mu_f \mu_{\tilde{A}}| \cdot |\mu_C \mu_{\tilde{C}} (\mu_{\mathcal{D}} - \mu_B \mu_C + \mu_B \mu_C) (\mu_{\tilde{\mathcal{D}}} - \mu_{\tilde{B}} \mu_{\tilde{C}} + \mu_{\tilde{B}} \mu_{\tilde{C}}) - \mu_B \mu_{\tilde{B}}| \\
&\leq |\mu_A \mu_f \mu_{\tilde{A}}| \cdot [|\mu_C \mu_{\tilde{C}} \mu_B \mu_C \mu_{\tilde{B}} \mu_{\tilde{C}} - \mu_B \mu_{\tilde{B}}| + |\mu_C \mu_{\tilde{C}} (\mu_{\mathcal{D}} - \mu_B \mu_C) (\mu_{\mathcal{D}} - \mu_B \mu_C)| \\
&\quad + |\mu_C \mu_{\tilde{C}} \mu_{\tilde{B}} \mu_{\tilde{C}} (\mu_{\mathcal{D}} - \mu_B \mu_C)| + |\mu_C \mu_{\tilde{C}} \mu_B \mu_C (\mu_{\tilde{\mathcal{D}}} - \mu_{\tilde{B}} \mu_{\tilde{C}})|] \\
&\stackrel{(a)}{\leq} |\mu_f \mu_A \mu_{\tilde{A}} \mu_B \mu_{\tilde{B}}| |\mu_{\tilde{C}}^2 \mu_C^2 - 1| + |\mu_f| \left(\Delta(k) \Delta(\tilde{k}) + \Delta(k) + \Delta(\tilde{k}) \right) \\
&\stackrel{(b)}{\leq} 8\epsilon + \frac{4\epsilon}{\sqrt{1 - \tanh \beta}} \left(\Delta(k) + \Delta(\tilde{k}) + \Delta(k) \Delta(\tilde{k}) \right) \\
&\stackrel{(c)}{\leq} 8\epsilon + 4\epsilon e^\beta (2\eta + \eta^2) \leq \eta.
\end{aligned}$$

Inequality (a) follows from (7.3). Inequality (b) uses Equation (7.2) and $|\mu_f| \leq \tau$. We showed that $\Delta(1) \leq \eta$. In (c) and using $\Delta(1) \leq \eta$, we use induction on the function Δ : the assumption $\epsilon < \min\{\frac{\eta}{16}, e^{-\beta}/24\}$ and $\Delta(k) < \eta$ for all $k < d$ gives $\Delta(d) \leq \eta$. \square

8. Sample complexity for exact structure learning problem

8.1. Samples necessary for structure learning (proof of Theorem 3.2)

We construct a family of trees that is difficult to distinguish. Suppose that p is odd (for simplicity) and let the graph T_0 be a path with associated parameters θ^0 given by $\theta_{i,i+1}^0 = \alpha$ for odd values of i and $\theta_{i,i+1}^0 = \beta$ for even values of i . For each odd value of $i \leq p-2$ we let θ^i be equal to θ^0 everywhere except $\theta_{i,i+1}^i = 0$ and $\theta_{i,i+2}^i = \alpha$. There are $(p+1)/2$ models in total (including θ^0).

The symmetrized KL-divergence between two zero-field Ising models θ and θ' has the convenient form

$$J(\theta \| \theta') \triangleq D(\theta \| \theta') + D(\theta' \| \theta) = \sum_{i < j} (\theta_{ij} - \theta'_{ij})(\mu_{ij} - \mu'_{ij}).$$

Here μ_{ij} and μ'_{ij} are the pairwise correlations between nodes i and j .

A small calculation leads to

$$\begin{aligned}
J(\theta^0 \| \theta^i) &= 2\alpha(\tanh \alpha - \tanh \alpha \tanh \beta) \\
&= 2\alpha(\tanh \alpha) \frac{2e^{-\beta}}{e^\beta + e^{-\beta}} \\
&\leq 4\alpha^2 e^{-2\beta}.
\end{aligned} \tag{8.1}$$

Here we used $\tanh \alpha \leq \alpha$ for $\alpha \geq 0$. Equation (8.1) can be plugged into Fano's inequality to complete the proof of Theorem 3.2.

We use the following version of Fano's inequality, which can be found, for example, as Corollary 2.6 in [30]. It gives a lower bound on the error probability in terms of the KL-divergence between pairs of points in the parameter space, where KL-divergence between two distributions P and Q on a space \mathcal{X} is defined as

$$D(P\|Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

Lemma 8.1 (Fano's inequality). *Assume that $M \geq 2$ and that Θ contains elements $\theta_0, \theta_1, \dots, \theta_M$. Let Q_{θ_j} denote the probability law of the observation X under model θ_j . If*

$$\frac{1}{M+1} \sum_{j=1}^M D(Q_{\theta_j} \| Q_{\theta_0}) \leq \gamma \log M \quad (8.2)$$

for $0 < \gamma < 1/8$, then the probability of error is bounded as $p_e \geq \frac{\log(M+1)-1}{\log M} - \gamma$.

8.2. Sufficient samples for structure learning (proof of Theorem 3.3)

Using Definition 5.4, Chow-Liu tree, T^{CL} recovers strong edges under the event $\mathsf{E}^{\text{strong}}(\epsilon)$. Thus, if the weakest edges in the original tree, T is strong enough, we can guarantee that with high probability, $\mathsf{T}^{\text{CL}} = \mathsf{T}$. Note that using lemma 9.9, the event $\mathsf{E}^{\text{strong}}(\epsilon)$ defined in Equation (5.4) occurs with probability $1 - \delta$ with $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$. Assuming that the original tree has parameters $\theta \in \Omega_{\alpha, \beta}$, the weakest edge should satisfy the following property to guarantee the correct recovery of all the edges of the tree T under the event $\mathsf{E}^{\text{strong}}(\epsilon)$:

$$\tanh \alpha \geq \frac{4\epsilon}{\sqrt{1 - \tanh \beta}}.$$

Thus the number of samples should satisfy:

$$n > \frac{16}{\tanh^2(\alpha)(1 - \tanh \beta)} \log \frac{2p^2}{\delta}.$$

Also $\tanh^{-1}(x) \leq 1/x + 1$ and $\frac{1}{1 - \tanh \beta} \leq e^\beta$. Hence $n > 6e^{2\beta} \max\{\alpha^{-2}, 1\} \log \frac{p}{\delta}$ guarantees $\mathsf{T}^{\text{CL}} = \mathsf{T}$ with probability at least $1 - \delta$.

Theorem 3.5. We assume that the tree we study is a Markov chain such that $x_1 \rightarrow x_2 \rightarrow \dots \rightarrow x_p$. We choose M different models over this Markov chain each denoted by the edge parameters $\theta^{(m)}$. We assume that for $i = 1, \dots, p-1$, we have $\theta_{i, i+1}^{(1)} = \text{atanh}(\eta)$. We also assume that in the m -th model we have

$\theta_{m,m+1}^{(m)} = 0$ and the remaining edge weights $\theta_{i,i+1}^{(m)} = \text{atanh}(\eta)$. Thus, the number of models we try to discriminate is $M = p$.

$$\begin{aligned} \max_{i,j} |\mathbb{E}_{\theta^{(m')}}[X_i X_j] - \mathbb{E}_{\theta^{(m)}}[X_i X_j]| &\geq \eta \\ J(\theta^m \parallel \theta^{m'}) &\leq 2\eta \text{atanh}(\eta) \end{aligned}$$

Using Fano's inequality in Lemma 8.1 gives the bound. \square

9. Control of events $\mathbf{E}^{\text{corr}}(\epsilon)$, $\mathbf{E}^{\text{strong}}(\epsilon)$, and $\mathbf{E}^{\text{cascade}}(\epsilon)$

We state a standard form of Hoeffding inequality:

Lemma 9.1 (Hoeffding's Inequality). *Let Z^1, \dots, Z^n be n i.i.d. random variables taking values in the interval $[-1, 1]$ and let $S_n = \frac{1}{n}(Z^1 + \dots + Z^n)$. Then, $\Pr[|S_n - \mathbb{E}Z| > t] \leq 2 \exp(-nt^2/2)$.*

Lemma 9.2. *The event $\mathbf{E}^{\text{corr}}(\epsilon)$ defined in (5.2) occurs with probability at least $1 - 2p^2 \exp(-n\epsilon^2/2)$.*

Proof. For each pair of nodes w, \tilde{w} , we define $Z = X_w X_{\tilde{w}}$ and Hoeffding's inequality stated in Lemma 9.1 to get $\Pr[|\mu_{w,\tilde{w}} - \hat{\mu}_{w,\tilde{w}}| > \epsilon] \leq 2 \exp(-n\epsilon^2/2)$. Union bound over pairs of nodes completes the proof. \square

We use the standard Bernstein's inequality quoted from [30].

Lemma 9.3 (Bernstein's Inequality). *Let Z^1, \dots, Z^n be i.i.d. random variables. Suppose that for all i , $|Z^i - \mathbb{E}[Z]| \leq M$ almost surely. Then, for all positive t ,*

$$\Pr \left[\left| \sum_{i=1}^n Z^i - n\mathbb{E}[Z] \right| \geq t \right] \leq 2 \exp \left(- \frac{t^2}{2n\text{VAR}(Z) + \frac{2}{3}Mt} \right)$$

For a given pair of nodes u, \tilde{u} and edge $e = (w, \tilde{w}) \in \text{path}_{\mathcal{T}}(u, \tilde{u})$, we define the random variable

$$Z_{e,u,\tilde{u}} = X_w X_{\tilde{w}} - X_u X_{\tilde{u}} = X_w X_{\tilde{w}} (1 - X_u X_w X_{\tilde{w}} X_{\tilde{u}}). \quad (9.1)$$

This random variable is used to study the event $e \notin \mathcal{T}^{\text{CL}}$ and $(u, \tilde{u}) \in \mathcal{T}^{\text{CL}}$ which happens only if $|\hat{\mu}_{u\tilde{u}}| \geq |\hat{\mu}_e|$. To bound the probability of this event we study the deviation bounds on the random variable $Z_{e,u,\tilde{u}}$. Let $\mathcal{A} = \text{path}_{\mathcal{T}}(u, \tilde{u}) \setminus \{e\}$ so that $\mu_{u\tilde{u}} = \mu_{\mathcal{A}} \mu_e$.

Lemma 9.4. *For all $u, \tilde{u} \in \mathcal{V}$ and $e = (w, \tilde{w}) \in \text{path}_{\mathcal{T}}(u, \tilde{u})$, given n samples $Z_{e,u,\tilde{u}}^{(1)}, Z_{e,u,\tilde{u}}^{(2)}, \dots, Z_{e,u,\tilde{u}}^{(n)}$ defined in (9.1), with probability at least $1 - \delta/2$*

$$\left| \sum_{i=1}^n Z_{e,u,\tilde{u}}^{(i)} - n\mu_e(1 - \mu_{\mathcal{A}}) \right| \leq \max \left\{ 16n\epsilon^2, 4n\epsilon\sqrt{1 - \mu_{\mathcal{A}}} \right\}, \quad (9.2)$$

where $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$ and $\mathcal{A} = \text{path}_{\mathcal{T}}(u, \tilde{u}) \setminus \{e\}$.

Proof. It follows from the fact that P is Markov with respect to \mathbb{T} and $e = (w, \tilde{w}) \in \text{path}_{\mathbb{T}}(u, \tilde{u})$ that $X_w X_{\tilde{w}}$ and $X_u X_w X_{\tilde{w}} X_{\tilde{u}}$ are independent. Note that

$$P(X_w X_{\tilde{w}} = 1) = \frac{1 + \mu_e}{2}, \quad P(X_w X_{\tilde{w}} = -1) = \frac{1 - \mu_e}{2}.$$

Similarly, the distribution of $X_u X_w X_{\tilde{u}} X_{\tilde{w}}$ is function of $\mu_{\mathcal{A}}$. As a result, the random variable $Z_{e,u,\tilde{u}} \in \{-2, 0, 2\}$ defined in (9.1) has the following distribution:

$$Z = \begin{cases} -2 & \text{w.p. } \frac{1-\mu_e}{2} \frac{1-\mu_{\mathcal{A}}}{2} \\ 0 & \text{w.p. } \frac{1+\mu_{\mathcal{A}}}{2} \\ +2 & \text{w.p. } \frac{1+\mu_e}{2} \frac{1-\mu_{\mathcal{A}}}{2} \end{cases}.$$

The first and second moments of Z are $\mathbb{E}[Z_{e,u,\tilde{u}}] = \mu_e(1 - \mu_{\mathcal{A}})$ and $\text{Var}[Z_{e,u,\tilde{u}}] = (1 - \mu_{\mathcal{A}})[2 - \mu_e^2(1 - \mu_{\mathcal{A}})]$. By Bernstein's inequality (Lemma 9.3), with probability at least $1 - \delta/2$

$$\left| \sum_{i=1}^n Z_{e,u,\tilde{u}}^{(i)} - n\mathbb{E}[Z] \right| \leq n \max \left\{ \frac{8}{3n} \log\left(\frac{4}{\delta}\right), \sqrt{\frac{4\text{Var}[Z_{e,u,\tilde{u}}]}{n} \log\left(\frac{4}{\delta}\right)} \right\}.$$

Using a union bound for all pairs of nodes u, \tilde{u} and edges $e = (w, \tilde{w}) \in \text{path}_{\mathbb{T}}(u, \tilde{u})$,

$$\begin{aligned} & \left| \sum_{i=1}^n Z_{e,u,\tilde{u}}^{(i)} - n\mu_e(1 - \mu_{\mathcal{A}}) \right| \\ & \leq n \max \left\{ \frac{8}{3n} \log\left(\frac{4p^3}{\delta}\right), \sqrt{\frac{4(1 - \mu_{\mathcal{A}})[2 - \mu_e^2(1 - \mu_{\mathcal{A}})]}{n} \log\left(\frac{4p^3}{\delta}\right)} \right\}. \end{aligned}$$

$2 - \mu_e^2(1 - \mu_{\mathcal{A}}) \leq 2$ and definition of ϵ gives $\sqrt{\frac{8(1 - \mu_{\mathcal{A}})[2 - \mu_e^2(1 - \mu_{\mathcal{A}})]}{n} \log\left(\frac{4p^3}{\delta}\right)} \leq 4\epsilon\sqrt{1 - \mu_{\mathcal{A}}}$. Also, $\frac{8}{3n} \log\left(\frac{4p^3}{\delta}\right) \leq 4\epsilon^2$ which gives the lemma. \square

Similarly to the above discussion, let $Y_{e,u,\tilde{u}} = X_w X_{\tilde{w}} + X_u X_{\tilde{u}}$. The proof of the following lemma follows the same steps as Lemma 9.4.

Lemma 9.5. *Given n samples, let $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$. Then, with probability at least $1 - \delta$, for all $u, \tilde{u} \in \mathcal{V}$ and $e = (w, \tilde{w}) \in \text{path}_{\mathbb{T}}(u, \tilde{u})$, and $Y_{e,u,\tilde{u}}^{(1)}, Y_{e,u,\tilde{u}}^{(2)}, \dots, Y_{e,u,\tilde{u}}^{(n)}$ defined in (9.1),*

$$\left| \sum_{i=1}^n Y_{e,u,\tilde{u}}^{(i)} - n\mu_e(1 + \mu_{\mathcal{A}}) \right| \leq \max \left\{ 16n\epsilon^2, 4n\epsilon\sqrt{(1 + \mu_{\mathcal{A}})} \right\}. \quad (9.3)$$

We defined the event $\mathbf{E}^{\text{strong}}(\epsilon)$ in (5.4) as the event that all the strong edges in \mathbb{T} as defined in (5.3) are recovered in \mathbb{T}^{CL} . In Lemma 9.8 we show that the deviation bounds for the variables $Z_{e,u,\tilde{u}}$ and $Y_{e,u,\tilde{u}}$ stated in (9.2) and (9.3) imply that $\mathbf{E}^{\text{strong}}(\epsilon)$ holds.

The following lemma is a well-known consequence of max-weight spanning tree algorithm used to construct the Chow-Liu tree.

Lemma 9.6 (Error characterization in Chow-Liu tree). *Consider the complete graph on p nodes with weights $|\hat{\mu}_{ij}|$ on each edge (i, j) . Let \mathbf{T}^{CL} be the maximum weight spanning tree of this graph. If edge $(w, \tilde{w}) \notin \mathbf{T}^{\text{CL}}$, then $|\hat{\mu}_{w\tilde{w}}| \leq |\hat{\mu}_{ij}|$ for all $(i, j) \in \text{path}_{\mathbf{T}^{\text{CL}}}(w, \tilde{w})$.*

Proof. We include a proof for completeness. For edge $(w, \tilde{w}) \notin \mathbf{T}^{\text{CL}}$, if there is an edge $(i, j) \in \text{path}_{\mathbf{T}^{\text{CL}}}(w, \tilde{w})$ such that $|\hat{\mu}_{w\tilde{w}}| \geq |\hat{\mu}_{ij}|$, then \mathbf{T}^{CL} cannot be the maximum weight spanning tree. To show that, we look at the tree \mathbf{T}' such that $(w, \tilde{w}) \in \mathbf{T}'$ and $(i, j) \notin \mathbf{T}'$ and all other edges of \mathbf{T}^{CL} appear in \mathbf{T}' . Note that: $\text{weight}(\mathbf{T}') \triangleq \sum_{e \in \mathbf{T}'} |\hat{\mu}_e| = \sum_{e \in \mathbf{T}^{\text{CL}}} |\hat{\mu}_e| + |\hat{\mu}_{w\tilde{w}}| - \hat{\mu}_{ij} \geq \text{weight}(\mathbf{T}^{\text{CL}})$. \square

Lemma 9.7. *If there exists a pair of edges $f = (w, \tilde{w})$ and $g = (u, \tilde{u})$ such that $f \in \mathbf{T}$, $f \notin \mathbf{T}^{\text{CL}}$, $g \in \mathbf{T}^{\text{CL}}$, $g \notin \mathbf{T}$ and additionally $f \in \text{path}_{\mathbf{T}}(u, \tilde{u})$ and $g \in \text{path}_{\mathbf{T}^{\text{CL}}}(w, \tilde{w})$. Then,*

$$\left(\sum_{i=1}^n Z_{f,u,\tilde{u}}^{(i)} \right) \left(\sum_{i=1}^n Y_{f,u,\tilde{u}}^{(i)} \right) < 0$$

Proof. Using Lemma 9.6, $f = (w, \tilde{w}) \notin \mathbf{T}^{\text{CL}}$ and $g = (u, \tilde{u}) \in \text{path}_{\mathbf{T}^{\text{CL}}}(w, \tilde{w})$ gives $|\hat{\mu}_f| \leq |\hat{\mu}_g|$. Hence $|\hat{\mu}_f|^2 \leq |\hat{\mu}_g|^2$ and

$$\begin{aligned} 0 &\geq |\hat{\mu}_f|^2 - |\hat{\mu}_g|^2 = (\hat{\mu}_f - \hat{\mu}_g)(\hat{\mu}_f + \hat{\mu}_g) \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n X_w^{(i)} X_{\tilde{w}}^{(i)} - X_u^{(i)} X_{\tilde{u}}^{(i)} \right) \left(\sum_{i=1}^n X_w^{(i)} X_{\tilde{w}}^{(i)} + X_u^{(i)} X_{\tilde{u}}^{(i)} \right) \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n Z_{f,u,\tilde{u}}^{(i)} \right) \left(\sum_{i=1}^n Y_{f,u,\tilde{u}}^{(i)} \right) \end{aligned}$$

where in the last equality we used $f \in \text{path}_{\mathbf{T}}(u, \tilde{u})$ and the definition of random variables $Z_{f,u,\tilde{u}}$ and $Y_{f,u,\tilde{u}}$. \square

Lemma 9.8. *Under the event described in Lemmas 9.4 and 9.5, if there is an edge $f \in \mathbf{T}$ missing from the Chow-Liu tree, $f \notin \mathbf{T}^{\text{CL}}$, then $|\mu_f| \leq \tau = \frac{4\epsilon}{\sqrt{1 - \tanh \beta}}$.*

Proof. We use Lemma 10.1 for $(w, \tilde{w}) = f$ to show that for the edge $f \in \mathbf{T}$ and $f \notin \mathbf{T}^{\text{CL}}$, there exists an edge $g = (u, \tilde{u}) \in \mathbf{T}^{\text{CL}}$ such that $g \notin \mathbf{T}$, $f \in \text{path}_{\mathbf{T}}(u, \tilde{u})$ and $g \in \text{path}_{\mathbf{T}^{\text{CL}}}(w, \tilde{w})$ (Figure 3). Then according to Lemma 9.7, $\left(\sum_{i=1}^n Z_{f,u,\tilde{u}}^{(i)} \right) \left(\sum_{i=1}^n Y_{f,u,\tilde{u}}^{(i)} \right) < 0$. Note that $\mathbb{E}Z_{f,u,\tilde{u}} = \mu_f(1 - \mu_{\mathcal{A}})$ and $\mathbb{E}Y_{f,u,\tilde{u}} = \mu_f(1 + \mu_{\mathcal{A}})$ using the definition $\mathcal{A} = \text{path}_{\mathbf{T}}(u, \tilde{u}) \setminus \{f\}$. Hence $\mathbb{E}Z_{f,u,\tilde{u}} \mathbb{E}Y_{f,u,\tilde{u}} \geq 0$. Thus, to satisfy $\left(\sum_{i=1}^n Z_{f,u,\tilde{u}}^{(i)} \right) \left(\sum_{i=1}^n Y_{f,u,\tilde{u}}^{(i)} \right) < 0$, either one of the following inequalities must hold:

$$\begin{aligned} \left| \sum_{i=1}^n Z_{f,u,\tilde{u}}^{(i)} - n\mathbb{E}Z_{f,u,\tilde{u}} \right| &\geq n|\mathbb{E}Z_{f,u,\tilde{u}}| \quad \text{or} \\ \left| \sum_{i=1}^n Y_{f,u,\tilde{u}}^{(i)} - n\mathbb{E}Y_{f,u,\tilde{u}} \right| &\geq n|\mathbb{E}Y_{f,u,\tilde{u}}|. \end{aligned}$$

Under the events described in Lemmas 9.4 and 9.5, this holds only if either one of these inequalities holds:

$$\begin{aligned} |\mu_f(1 - \mu_{\mathcal{A}})| &\leq \max \left\{ 4\epsilon^2, 4\epsilon\sqrt{(1 - \mu_{\mathcal{A}})} \right\} \\ |\mu_f(1 + \mu_{\mathcal{A}})| &\leq \max \left\{ 4\epsilon^2, 4\epsilon\sqrt{(1 + \mu_{\mathcal{A}})} \right\}. \end{aligned}$$

This implies that

$$\begin{aligned} |\mu_f| &\leq \max \left\{ \frac{4\epsilon}{\sqrt{1 - \mu_{\mathcal{A}}}}, \frac{4\epsilon^2}{1 - \mu_{\mathcal{A}}}, \frac{4\epsilon}{\sqrt{1 + \mu_{\mathcal{A}}}}, \frac{16\epsilon^2}{1 + \mu_{\mathcal{A}}} \right\} \\ &\leq \max \left\{ \frac{4\epsilon}{\sqrt{1 - \tanh \beta}}, \frac{16\epsilon^2}{1 - \tanh \beta} \right\}. \end{aligned}$$

Note that if $\frac{4\epsilon}{\sqrt{1 - \tanh \beta}} \geq 1$ then the bound on $|\mu_f| \leq 1$ is trivial. And if $\frac{4\epsilon}{\sqrt{1 - \tanh \beta}} \leq 1$, then $\frac{4\epsilon}{\sqrt{1 - \tanh \beta}} > \frac{16\epsilon^2}{1 - \tanh \beta}$. Hence, $|\mu_f| \leq \frac{4\epsilon}{\sqrt{1 - \tanh \beta}}$. \square

Lemma 9.9. *With $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$, event $\mathbf{E}^{\text{strong}}(\epsilon)$ defined in (5.4) happens with probability at least $1 - \delta$.*

Proof. Lemma 9.8 shows that, under the event described in Lemmas 9.4 and 9.5 the missing edges $f \in \mathbf{T}$ and $f \notin \mathbf{T}^{\text{CL}}$ satisfy $|\mu_f| \leq \tau$. Hence, Lemmas 9.4 and 9.5 show that with probability at least $1 - \delta$, all edges $e \in \mathbf{T}$ such that $|\mu_e| \geq \tau$ are recovered by Chow-Liu algorithm $e \in \mathbf{T}^{\text{CL}}$. \square

Lemma 9.10. *Let the distribution $P(x) \in \mathcal{P}(\mathbf{T})$ come from the class of zero-field Ising models on the tree $\mathbf{T} = (\mathcal{V}, \mathcal{E})$. For all $e = (i, j) \in \mathcal{E}$ let $Y_e = X_i X_j$. Then the set of random variables Y_e are jointly independent.*

This lemma follows from the factorization of distribution $P(x) \in \mathcal{P}(\mathbf{T})$ in Equation (1.1).

The following lemma is restatement of Theorem 9 in [4] for our desired application.

Lemma 9.11. *[[4], Theorem 9] Let $Y_1, \dots, Y_d \in \{-1, +1\}^d$ be d independent random variables with $\mathbb{E}[Y_j] = \mu_j$. Having n i.i.d. samples $y_j^{(i)}$ for $j = 1, \dots, d$ and $i = 1, \dots, n$, let $\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n y_j^{(i)}$ be the empirical average of Y_j . Then for any $\gamma > 0$,*

$$\Pr \left[\left| \prod_{j=1}^d \hat{\mu}_j - \prod_{j=1}^d \mu_j \right| \geq \gamma \right] \leq \frac{8}{\gamma} \exp(-\gamma^2 n / 32).$$

Proof.

$$\prod_{j=1}^d \hat{\mu}_j = \prod_{j=1}^d \left(\frac{1}{n} \sum_{i=1}^n y_j^{(i)} \right) = \frac{1}{n^d} \prod_{j=1}^d \sum_{i_j=1}^n y_j^{(i_j)} = \frac{1}{n^d} \sum_{i_1, \dots, i_d=1}^n \prod_{j=1}^d y_j^{(i_j)}$$

For every $l \in \{1, \dots, n\}$, one can write

$$\prod_{j=1}^d \hat{\mu}_j = \frac{1}{n^d} \sum_{i_1, \dots, i_d=1}^n \prod_{j=1}^d y_j^{(i_j+l)}$$

where $i_j + l$ are taken modulo n . When we take average over all l , we have

$$\prod_{j=1}^d \hat{\mu}_j = \frac{1}{n^d} \sum_{i_1, \dots, i_d=1}^n \frac{1}{n} \sum_{l=1}^n \prod_{j=1}^d y_j^{(i_j+l)} \quad (9.4)$$

For every initial multi-indices $i_* = (i_1, \dots, i_d)$, we define

$$d(i_*) = \begin{cases} 0 & \text{if } \left| \frac{1}{n} \sum_{l=1}^n \prod_{j=1}^d y_j^{(i_*+l)} - \prod_{j=1}^d \mu_j \right| \geq \gamma/2 \\ 2 & \text{otherwise.} \end{cases}$$

By definition of $d(i_*)$, we have

$$\left| \frac{1}{n} \sum_{l=1}^n \prod_{j=1}^d y_j^{(i_*+l)} - \prod_{j=1}^d \mu_j \right| \leq d(i_*) + \gamma/2 \quad (9.5)$$

Also, note that for any fixed i_* , the random variables $\prod_{j=1}^d y_j^{(i_*+1)}, \dots, \prod_{j=1}^d y_j^{(i_*+n)}$ are independent with expectation $\prod_{j=1}^d \mu_j$. Therefore, using Hoeffding inequality 9.1, we have

$$\mathbb{E}[d(i_*)] \leq 4 \exp(-\gamma^2 n/32). \quad (9.6)$$

We use Equations (9.4) and (9.5) to get the bound:

$$\begin{aligned} \Pr \left[\left| \prod_{j=1}^d \hat{\mu}_j - \prod_{j=1}^d \mu_j \right| \geq \gamma \right] &= \Pr \left[\left| \frac{1}{n^d} \sum_{i_1, \dots, i_d=1}^n \frac{1}{n} \sum_{l=1}^n \prod_{j=1}^d y_j^{(i_j+l)} - \prod_{j=1}^d \mu_j \right| \geq \gamma \right] \\ &\leq \Pr \left[\frac{1}{n^d} \sum_{i_1, \dots, i_d=1}^n \left| \frac{1}{n} \sum_{l=1}^n \prod_{j=1}^d y_j^{(i_j+l)} - \prod_{j=1}^d \mu_j \right| \geq \gamma \right] \\ &\leq \Pr \left[\frac{1}{n^d} \sum_{i_1, \dots, i_d=1}^n d(i_*) \geq \gamma/2 \right] \\ &\leq \frac{8}{\gamma} \exp(-\gamma^2 n/32), \end{aligned}$$

where in the last inequality we used Markov inequality and Equation (9.6). \square

Lemma 9.12. *The event $E^{\text{cascade}}(\epsilon)$ defined in (5.5) occurs with probability at least $1 - \frac{4p^2}{\epsilon} \exp(-\epsilon^2 n/32)$.*

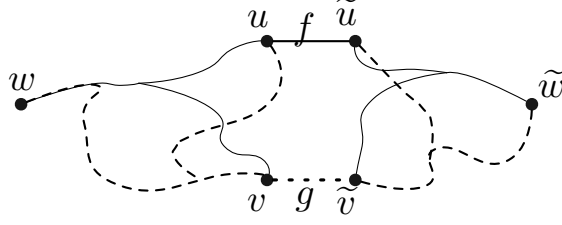


FIG 3. The schematics showing the pairs of edges f and g satisfying properties given in lemma 10.1. The solid lines represent the tree T and the dashed lines represent the tree \hat{T} . There always exists pairs of edges f and g such that $f = (u, \tilde{u}) \in \text{path}_T(w, \tilde{w}) \cap \text{path}_T(v, \tilde{v})$ and $g = (v, \tilde{v}) \in \text{path}_{\hat{T}}(w, \tilde{w}) \cap \text{path}_{\hat{T}}(u, \tilde{u})$.

Proof. For given pair of nodes w, \tilde{w} , define $\text{path}_T(w, \tilde{w}) = \{e_1, e_2, \dots, e_d\}$ and use the notation $\mu_i = \mu_{e_i}$ and $\hat{\mu}_i = \hat{\mu}_{e_i}$. We apply Lemma 9.11 to show

$$\Pr\left[\left|\prod_{j=1}^d \hat{\mu}_j - \prod_{j=1}^d \mu_j\right| \leq \epsilon\right] \geq \frac{8}{\epsilon} \exp(-\epsilon^2 n / 32).$$

Applying a union bound over $\binom{p}{2} \leq p^2/2$ pairs of nodes w, \tilde{w} provides a lower bound for probability of $\mathbf{E}^{\text{cascade}}(\epsilon)$. \square

10. Two trees lemma

Lemma 10.1. *Let T and \hat{T} be two spanning trees on set of nodes \mathcal{V} . Let w, \tilde{w} be a pair of nodes such that $\text{path}_T(w, \tilde{w}) \neq \text{path}_{\hat{T}}(w, \tilde{w})$. Then there exists a pair of edges $f \triangleq (u, \tilde{u}) \in \text{path}_T(w, \tilde{w})$ and $g \triangleq (v, \tilde{v}) \in \text{path}_{\hat{T}}(w, \tilde{w})$ such that*

- (i) $f \notin \text{path}_{\hat{T}}(w, \tilde{w})$ and $g \notin \text{path}_T(w, \tilde{w})$
- (ii) $f \in \text{path}_T(v, \tilde{v})$ and $g \in \text{path}_{\hat{T}}(u, \tilde{u})$

Since $f \in \text{path}_T(w, \tilde{w})$ and $f \in \text{path}_T(v, \tilde{v})$, we can choose the labeling of the end points of the edges $f = (u, \tilde{u})$ and $g = (v, \tilde{v})$ such that $u, v \in \text{SubTree}_{T,f}(w)$ and $\tilde{u}, \tilde{v} \in \text{SubTree}_{T,f}(\tilde{w})$. [as in Figure 3]

Proof. Define a graph G such that the nodes on this graph are the nodes in $\text{path}_T(w, \tilde{w}) \cup \text{path}_{\hat{T}}(w, \tilde{w})$. We define two types of edges in this graph corresponding to the edges in $\text{path}_T(w, \tilde{w})$ and the edges in $\text{path}_{\hat{T}}(w, \tilde{w})$. Then, we merge any pair of nodes between which there are both types of edges. Note that since $\text{path}_T(w, \tilde{w}) \neq \text{path}_{\hat{T}}(w, \tilde{w})$, the nodes w and \tilde{w} will not be merged with each other during this process. The remaining nodes in the $\text{path}_T(w, \tilde{w})$ are labeled u_1, u_2, \dots, u_{K-1} and the remaining edges are labeled f_1, f_2, \dots, f_K . Similarly, the remaining nodes in $\text{path}_{\hat{T}}(w, \tilde{w})$ are labeled v_1, v_2, \dots, v_{L-1} and the remaining edges are labeled g_1, g_2, \dots, g_L . Where K and L are positive integers. We label these edges so that $f_k = (u_{k-1}, u_k)$ and $g_l = (v_{l-1}, v_l)$. This

notation is consistent if we define $u_0 = v_0 = w$ and $u_K = v_L = \tilde{w}$. Note that in this notation one node may have several labels.

According to the definition, $f_k \in \text{path}_T(w, \tilde{w})$ and $g_l \in \text{path}_{\hat{T}}(w, \tilde{w})$ for all $k = 1, \dots, K$ and $l = 1, \dots, L$. The process of merging nodes as described guarantees that $f_k \notin \text{path}_{\hat{T}}(w, \tilde{w})$ and $g_l \notin \text{path}_T(w, \tilde{w})$. Thus, in order to prove the existence of edges with the properties given in the lemma, all we have to do is to find a pair of edges f_{k^*} and g_{l^*} such that $f_{k^*} \in \text{path}_T(v_{l^*-1}, v_{l^*})$ and $g_{l^*} \in \text{path}_{\hat{T}}(u_{k^*-1}, u_{k^*})$.

Now we define another graph \hat{G} to be the graph with the nodes $\hat{\mathcal{V}} = \{w, \tilde{w}, u_1, u_2, \dots, u_K, v_1, v_2, \dots, v_L\}$. There are two types of edges in this graph corresponding to the connectivity in the spanning tree T and \hat{T} . We denote the edges corresponding to the tree T by \mathcal{E} and the edges corresponding to the tree \hat{T} by $\hat{\mathcal{E}}$. Consequently, \mathcal{E} and $\hat{\mathcal{E}}$ both construct two different spanning trees of the graph \hat{G} on the nodes \mathcal{V} .

The set of edges \mathcal{E} determine a spanning tree denoted by S and satisfy all the conditional independence properties satisfied by the original spanning tree T on the nodes $\hat{\mathcal{V}}$. Concretely, any pair of nodes (i, j) in $\hat{\mathcal{V}}$ are connected by an edge in \mathcal{E} if and only if all the nodes in $\text{path}_T(i, j)$ have been removed from \mathcal{V} during construction of \hat{G} , i.e., $\hat{\mathcal{V}} \cap \text{path}_T(i, j) = \{i, j\}$.

Similarly, the edges $\hat{\mathcal{E}}$ are chosen so that the spanning tree implied them, denoted by \hat{S} satisfies all the conditional independence properties satisfied by the original spanning tree \hat{T} .

We define the set $\mathcal{D}_k = \{l; g_l \in \text{path}_{\hat{S}}(w, \tilde{w}) \cap \text{path}_S(u_{k-1}, u_k)\}$ to be the set of edges in $\text{path}_{\hat{S}}(w, \tilde{w})$ which are also in the path connecting the two ends of f_k in the tree \hat{S} . Similarly, for $l = 1, \dots, L$ we define the sets $\mathcal{C}_l = \{k; f_k \in \text{path}_S(w, \tilde{w}) \cap \text{path}_S(v_{l-1}, v_l)\}$. If there exists the pair integers k^* and l^* such that $l^* \in \mathcal{D}_{k^*}$ and $k^* \in \mathcal{C}_{l^*}$, then they correspond to the edges $f_{k^*} \in \text{path}_T(v_{l^*-1}, v_{l^*})$ and $g_{l^*} \in \text{path}_{\hat{T}}(u_{k^*-1}, u_{k^*})$. As stated previously, this pair of edges satisfy all the properties given in the lemma.

To find such pair of l and k we prove a “no gap” property of the sets $\{\mathcal{C}_l\}$, with a similar statement holding for $\{\mathcal{D}_k\}$. We will use the following notation in the statement of the “no gap” lemma. Let’s define

$$\underline{\mathcal{D}}_{k'} = \bigcup_{k=1}^{k'} \mathcal{D}_k, \quad \overline{\mathcal{D}}_{k'} = \bigcup_{k=K-k'+1}^K \mathcal{D}_k, \quad \underline{\mathcal{C}}_{l'} = \bigcup_{l=1}^{l'} \mathcal{C}_l, \quad \overline{\mathcal{C}}_{l'} = \bigcup_{l=L-l'+1}^L \mathcal{C}_l.$$

Lemma 10.2 (“No gap” lemma). *For any $l' > 0$, if $\underline{\mathcal{C}}_{l'} \neq \emptyset$, there exists $k' > 0$ such that $\underline{\mathcal{D}}_{k'} = \{1, 2, \dots, k'\}$. For any $l' > 0$, if $\overline{\mathcal{C}}_{l'} \neq \emptyset$, then there exists $k' \geq 0$ such that $\overline{\mathcal{D}}_{k'} = \{K - k', \dots, K\}$.*

To prove that the “no gap” property of the sets $\{\mathcal{C}_l\}$ and $\{\mathcal{D}_k\}$, we define the following function $M : \{0, 1, 2, \dots, K\} \rightarrow \{-1, 0, 1, 2, \dots, K\}$ such that $M(0) = -1$ and for $1 \leq k' \leq K$,

$$M(k') = \max\{i; i \in \bigcup_{l \in \underline{\mathcal{D}}_{k'}} \mathcal{C}_l\}. \quad (10.1)$$

Note that $\mathcal{D}_{k'} \subseteq \mathcal{D}_{k'+1}$ and consequently $\cup_{l \in \mathcal{D}_{k'}} \mathcal{C}_l \subseteq \cup_{l \in \mathcal{D}_{k'+1}} \mathcal{C}_l$. Thus, the function $M(k')$ is non-decreasing over its domain.

Also, due to the no gap property of the sequence of sets $\{\mathcal{C}_l\}$ and $\{\mathcal{D}_k\}$, we know that $\mathcal{D}_K = \{1, \dots, L\}$ and $\cup_{l=1}^L \mathcal{C}_l = \mathcal{C}_L = \{1, \dots, K\}$. Thus, $M(K) = K$. Define

$$k^* = \min\{k : 1 \leq k \leq K \text{ and } k \leq M(k)\}.$$

Thus, we know that $k^* \leq M(k^*)$ and $M(k^* - 1) < k^* - 1 < k^*$. So, $M(k^* - 1) < k^* \leq M(k^*)$. According to the definition given by equation (10.1), there exists some $l^* \in \mathcal{D}_{k^*}$ such that $k^* \in \mathcal{C}_{l^*}$.

As mentioned before, this pair integers determine the edges that satisfy the properties given in the lemma. \square

Proof of Lemma 10.2. For each node v_l we define a pointer $p_v(l) = \max\{k; u_k \in \text{path}_S(w, v_l)\}$. Note that $p_v(0) = 0$ and $p_v(L) = K$. Similarly we define the sequence of pointers $p_u(k) = \max\{l; v_l \in \text{path}_S(w, u_k)\}$.

These pointers determine the sequence of sets \mathcal{C}_l 's (and similarly for \mathcal{D}_k 's) as follows: For any $l > 0$, if $p_v(l) > p_v(l-1)$, $\mathcal{C}_l = \{p_v(l-1)+1, \dots, p_v(l)\}$. If $p_v(l) = p_v(l-1)$, then $\mathcal{C}_l = \emptyset$. And if $p_v(l) < p_v(l-1)$, then $\mathcal{C}_l = \{p_v(l)+1, \dots, p_v(l-1)\}$.

Based on this characterization of the sets \mathcal{C}_l 's, the nested version of the sets \mathcal{C}_l 's is

$$\mathcal{C}_{l'} = \left\{1, \dots, \max_{l \leq l'} p_v(l)\right\}.$$

This is nonempty if and only if $\max_{l \leq l'} p_v(l) > 0$. Similarly, we have that

$$\overline{\mathcal{C}}_{l'} = \left\{\min_{l \geq L-l'} p_v(l), \dots, K\right\}.$$

Similar reasoning gives

$$\mathcal{D}_{k'} = \left\{1, \dots, \max_{k \leq k'} p_u(k)\right\}$$

and

$$\overline{\mathcal{D}}_{k'} = \left\{\min_{k \geq K-k'} p_u(k), \dots, L\right\}.$$

Observe that $\max_{l \leq l'} p_v(l)$ as a function of $0 \leq l' \leq L$, is monotonically increasing, where $p_v(0) = 0$ and $p_v(L) = K$. Similarly, $\max_{k \leq k'} p_u(k)$ as a function of $0 \leq k' \leq K$, is monotonically increasing, where $p_u(0) = 0$ and $p_u(K) = L$. Consequently, there exists a pair of integers l and k such that $k \in \mathcal{C}_l$ and $l \in \mathcal{D}_k$. \square

Lemma 10.3. *Suppose that the distribution $P(x)$ is from an Ising model on tree \mathbb{T} with pairwise correlations $\mu_{w, \tilde{w}}$ for each $w, \tilde{w} \in \mathcal{V}$. Assume event $\mathbf{E}^{\text{corr}}(\epsilon)$ holds. Let $\hat{\mathbb{T}}$ be the Chow-Liu tree. For any pair of edges $(u, \tilde{u}) \in \text{path}_{\mathbb{T}}(v, \tilde{v})$ and $(v, \tilde{v}) \in \text{path}_{\hat{\mathbb{T}}}(u, \tilde{u})$ with $(u, \tilde{u}) \notin \hat{\mathbb{T}}$ and $(v, \tilde{v}) \notin \mathbb{T}$, we have*

$$|\mu_{u, \tilde{u}}| - 4\epsilon \leq |\mu_{v, \tilde{v}}| \leq |\mu_{u, \tilde{u}}|.$$

Proof. Lemma 9.6 applied to $(u, \tilde{u}) \notin \hat{\mathbf{T}}$ and $(v, \tilde{v}) \in \text{path}_{\hat{\mathbf{T}}}(u, \tilde{u})$ gives $|\hat{\mu}_{v, \tilde{v}}| \geq |\hat{\mu}_{u, \tilde{u}}|$. The first inequality now follows from the fact that on the event $\mathbf{E}^{\text{corr}}(\epsilon)$ we have $|\hat{\mu}_{u, \tilde{u}} - \mu_{u, \tilde{u}}| \leq 2\epsilon$ and $|\hat{\mu}_{v, \tilde{v}} - \mu_{v, \tilde{v}}| \leq 2\epsilon$.

The second inequality follows as $|\mu_{v, \tilde{v}}| = |\prod_{e \in \text{path}_T(v, \tilde{v})} \mu_e| \leq |\mu_{u, \tilde{u}}|$, since $\mu_{u, \tilde{u}}$ appears in the product and $|\mu_e| \leq 1$ for all e . \square

11. Information-projection lemmas

Lemma 11.1. *For a given tree $T = (\mathcal{V}, \mathcal{E})$, the projection of the distribution $P(x)$ onto the class of the Ising models on \mathbf{T} with no external field (as in Equation (1.1)) $\tilde{P}(x) = \Pi_{\mathbf{T}}(P) \triangleq \arg \min_{Q \in \mathcal{P}_{\mathbf{T}}} D(P\|Q)$ with the parameter vec-*

tors $\tilde{\theta} \in \Omega_{0, \infty}(T)$ has the following property: $\tanh(\tilde{\theta}_{ij}) = \mu_{ij}$ for all $(i, j) \in \mathcal{E}$, where $\mu_{ij} = \mathbb{E}_P X_i X_j$ is the pairwise correlation of the variables under the distribution $P(x)$. Also, we would have $D(P\|\tilde{P}) = -H(P) + \sum_{(i, j) \in \mathcal{E}} H_B(\frac{1+\mu_{ij}}{2})$ where $H_B(x) = x \log \frac{1}{x} + (1-x) \log \frac{1}{1-x}$ is the binary entropy function.

Proof. This lemma is a direct corollary of Theorem 3.3 in [11]. Note that $\tilde{P}(x)$ is the reverse I-projection of distribution $P(x)$ onto the exponential family with potential functions $X_i X_j$ for $(i, j) \in \mathcal{E}$. Hence $\mathbb{E}_{\tilde{P}} X_i X_j = \mathbb{E}_P X_i X_j$ for $(i, j) \in \mathcal{E}$ which gives $\tanh(\tilde{\theta}_{ij}) = \mu_{ij}$. Section 3.4.2 in [32] also addresses this problem in depth.

We provide the proof for the sake of completeness here: For the tree $\mathbf{T} = (\mathcal{V}, \mathcal{E})$ and any $Q \in \mathcal{P}_{\mathbf{T}}$ with parameter vector θ , using (1.1), Q could be factorized as $Q(x) = \prod_{(i, j) \in \mathcal{E}} (1 + \tanh \theta_{ij} x_i x_j)/2$, where $\mathbb{E}_Q[X_i X_j] = \tanh \theta_{ij}$. Let spin random variables $Y_e = X_i X_j$ for all $e = (i, j) \in \mathcal{E}$ so that $Q(Y_e = +) = (1 + \tanh \theta_{ij})/2$ and $P(Y_e = +) = (1 + \mu_{ij})/2$ where $\mu_{ij} = \mathbb{E}_P[X_i X_j]$.

Given the general distribution P and any $Q \in \mathcal{P}_{\mathbf{T}}$ with parameters θ ,

$$\begin{aligned} D(P\|Q) &= \mathbb{E}_P[\log \frac{P}{Q}] = -H(P) - \mathbb{E}_P[\log Q(X)] \\ &= -H(P) - \mathbb{E}_P[\log \prod_{(i, j) \in \mathcal{E}} \frac{1 + \tanh \theta_{ij} X_i X_j}{2}] \\ &= -H(P) - \sum_{e=(i, j) \in \mathcal{E}} \mathbb{E}_P[\log \frac{1 + \tanh \theta_{ij} Y_e}{2}] \\ &= -H(P) - \sum_{e \in \mathcal{E}} \mathbb{E}_P[\log Q(Y_e)] = \\ &= -H(P) + \sum_{e \in \mathcal{E}} H_B(P(Y_e)) + \sum_{e \in \mathcal{E}} D(P(Y_e)\|Q(Y_e)) \end{aligned}$$

$\tilde{P} = \arg \min_{Q \in \mathcal{P}_{\mathbf{T}}} D(P\|Q)$ with parameters $\tilde{\theta}$ satisfies the property $\tilde{P}(Y_e) = P(Y_e)$ for all $e \in \mathcal{E}$, hence $\tilde{\theta}_{ij} = \mu_{ij}$ for all $(i, j) \in \mathcal{E}$. Consequently, $D(P\|\tilde{P}) = -H(P) + \sum_{e \in \mathcal{E}} H_B(\frac{1+\mu_{ij}}{2})$. \square

Lemma 11.2. *Given distribution \hat{P} , the tree T^{CL} defined in Definition 3.1 can be found as the maximum weight spanning tree over a complete weighted graph where the weights of each edge (i, j) is $|\hat{\mu}_{ij}| = |\mathbb{E}_{\hat{P}}[X_i X_j]|$.*

Proof. Using Lemma 11.1,

$$\mathsf{T}^{\text{CL}} = \arg \min_{\mathsf{T} \in \mathcal{T}} \sum_{(i,j) \in \mathcal{E}_{\mathsf{T}}} H_B\left(\frac{1 + \hat{\mu}_{ij}}{2}\right) - H(\hat{P}) = \arg \min_{\mathsf{T} \in \mathcal{T}} \sum_{(i,j) \in \mathcal{E}_{\mathsf{T}}} H_B\left(\frac{1 + \hat{\mu}_{ij}}{2}\right).$$

The maximum weight spanning tree can be implemented greedily using Kruskal algorithm or Prim’s algorithm [10]. So, finding maximum weight spanning tree only depends on the order of the edges of the graph after being sorted.

$H_B(\frac{1 + \hat{\mu}_{ij}}{2})$ is a monotonically increasing function of $|\hat{\mu}_{ij}|$. So, sorting all the edges (i, j) in the complete graph based $H_B(\frac{1 + \hat{\mu}_{ij}}{2})$ or $|\hat{\mu}_{ij}|$ gives the same order. This gives

$$\mathsf{T}^{\text{CL}} = \arg \min_{\mathsf{T} \in \mathcal{T}} \sum_{(i,j) \in \mathcal{E}_{\mathsf{T}}} H_B\left(\frac{1 + \hat{\mu}_{ij}}{2}\right) = \arg \min_{\mathsf{T} \in \mathcal{T}} \sum_{(i,j) \in \mathcal{E}_{\mathsf{T}}} |\hat{\mu}_{ij}|. \quad \square$$

Discussion

Learning the structure of Ising models on trees with finite number of samples is proved to be impossible in the presence of extremely weak edges. Yet, we proved that using the maximum likelihood tree as an estimation of the structure would enable us to have an accurate estimation of the the original distribution for the purpose of subsequent inference tasks. The sample complexity of this estimation is given in this paper.

Acknowledgment

We thank Gregory Wornell, Lizhong Zheng, Gabor Lugosi, Devavrat Shah, David Gamarnik, Elchanan Mossel and Andrea Montanari for stimulating discussions.

References

- [1] ABBEEL, P., KOLLER, D., AND NG, A. Learning factor graphs in polynomial time and sample complexity. *JMLR* (2006).
- [2] AGARWALA, R., BAFNA, V., FARACH, M., PATERSON, M., AND THORUP, M. On the approximability of numerical taxonomy (fitting distances by tree metrics). *SIAM Journal on Computing* 28, 3 (1998), 1073–1085.
- [3] ANANDKUMAR, A., HUANG, F., HSU, D., AND KAKADE, S. Learning mixtures of tree graphical models. In *NIPS* (2012).
- [4] BABICHENKO, Y., BARMAN, S., AND PERETZ, R. Empirical distribution of equilibrium play and its testing application. *arXiv preprint arXiv:1310.7654* (2013).

- [5] BENTO, J., AND MONTANARI, A. Which graphical models are difficult to learn? In *NIPS* (2009).
- [6] BRESLER, G. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing* (2015), ACM, pp. 771–782.
- [7] BRESLER, G., MOSSEL, E., AND SLY, A. Reconstruction of Markov random fields from samples: Some observations and algorithms. In *APPROX* (2008).
- [8] CHOW, C., AND LIU, C. Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theory* 14, 3 (1968), 462–467.
- [9] CHOW, C.-Y., AND WAGNER, T. Consistency of an estimate of tree-dependent probability distributions (corresp.). *Information Theory, IEEE Transactions on* 19, 3 (1973), 369–371.
- [10] CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L., AND STEIN, C. *Introduction to algorithms*, vol. 6. MIT press Cambridge, 2001.
- [11] CSISZÁR, I., AND SHIELDS, P. C. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.
- [12] DASGUPTA, S. Learning polytrees. In *UAI* (1999).
- [13] DASKALAKIS, C., MOSSEL, M., AND ROCH, S. Phylogenies without branch bounds: Contracting the short, pruning the deep. *Research in Computational Molecular Biology: 13th Annual International Conference, RECOMB 2009, Tucson, AZ, USA, May 18-21, 2009. Proceedings* (2009), 451 – 465.
- [14] ERDÖS, P. L., STEEL, M. A., SZÉKELY, L., AND WARNOW, T. J. A few logs suffice to build (almost) all trees: Part ii. *Theoretical Computer Science* 221, 1 (1999), 77–118.
- [15] HEINEMANN, U., AND GLOBERSON, A. Infering with high girth graphical models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (2014), pp. 1260–1268.
- [16] JOG, V., AND LOH, P.-L. On model misspecification and KL separation for Gaussian graphical models. In *Information Theory (ISIT), 2015 IEEE International Symposium on* (2015), IEEE, pp. 1174–1178.
- [17] LAURITZEN, S. *Graphical models*. Oxford University Press, 1996.
- [18] LIU, H., XU, M., GU, H., GUPTA, A., LAFFERTY, J., AND WASSERMAN, L. Forest density estimation. *The Journal of Machine Learning Research* 12 (2011), 907–951.
- [19] MEILA, M., AND JORDAN, M. I. Learning with mixtures of trees. *The Journal of Machine Learning Research* 1 (2001), 1–48.
- [20] MEINSHAUSEN, N., AND BÜHLMANN, P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* (2006), 1436–1462.
- [21] MOSSEL, E. Distorted metrics on trees and phylogenetic forests. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4 (2007).
- [22] NARASIMHAN, M., AND BILMES, J. PAC-learning bounded tree-width graphical models. in *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference* (2004), 410–417.

- [23] PEARL, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
- [24] RAVIKUMAR, P., WAINWRIGHT, M., AND LAFFERTY, J. High-dimensional Ising model selection using ℓ_1 -regularized logistic regression. *Annals of Statistics* 38, 3 (2010), 1287–1319.
- [25] SANTHANAM, N. P., AND WAINWRIGHT, M. J. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. on Info. Theory* 58, 7 (2012), 4117–4134.
- [26] SREBRO, N. Maximum likelihood bounded tree-width Markov networks. In *UAI* (2001).
- [27] TAN, V. Y., ANANDKUMAR, A., TONG, L., AND WILLSKY, A. S. A large-deviation analysis of the maximum-likelihood learning of Markov tree structures. *Information Theory, IEEE Transactions on* 57, 3 (2011), 1714–1735.
- [28] TAN, V. Y., ANANDKUMAR, A., AND WILLSKY, A. S. Learning high-dimensional Markov forest distributions: Analysis of error rates. *The Journal of Machine Learning Research* 12 (2011), 1617–1653.
- [29] TANDON, R., SHANMUGAM, K., RAVIKUMAR, P. K., AND DIMAKIS, A. G. On the information theoretic limits of learning Ising models. In *Advances in Neural Information Processing Systems* (2014), pp. 2303–2311.
- [30] TSYBAKOV, A. B. *Introduction to nonparametric estimation*. Springer, 2008.
- [31] WAINWRIGHT, M. Estimating the “wrong” graphical model: Benefits in the computation-limited setting. *The Journal of Machine Learning Research* 7 (2006), 1829–1857.
- [32] WAINWRIGHT, M., AND JORDAN, M. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* 1, 1-2 (2008), 1–305.
- [33] WAINWRIGHT, M. J. Tree-reweighted belief propagation algorithms and approximate ML estimation via pseudo-moment matching. *IN AISTATS* (2003).
- [34] WU, R., SRIKANT, R., AND NI, J. Learning loosely connected Markov random fields. *Stochastic Systems* 3, 2 (2013), 362–404.